B.N. PSHENICHNYI

MÉTHODES NUMÉRIQUES DANS LES PROBLÈMES D'EXTRÉMUM

B. PCHÉNITCHNY ET Y. DANILINE

MÉTHODES NUMÉRIQUES DANS LES PROBLÈMES D'EXTRÉMUM

ÉDITIONS MIR • MOSCOU

PRÉFACE

Les méthodes numériques de résolution de divers problèmes d'extrémum ont pris ces dernières années un bel essor si bien que la bibliographie correspondante contient des centaines d'ouvrages. Cet intérêt n'est nullement fortuit, il reflète le rôle de premier plan que les problèmes d'extrémum jouent dans les applications. Le présent livre est précisément consacré à la recherche efficace du minimum d'une fonction dont les variables sont soumises à des contraintes.

Signalons tout de suite que les exigences formulées à l'égard de nouveaux algorithmes ne sont pas les mêmes qu'il y a dix ou quinze années où chaque nouvelle procédure de calcul pour tel ou tel problème de minimisation a été accueillie avec intérêt. Il ne suffit actuellement pas de construire un algorithme, il faut montrer de plus qu'il l'emporte sur ceux connus. On a donc à comparer l'efficacité de différents algorithmes, problème qui ne se résout pas, hélas, si facilement. En effet, on ne compare qu'en se basant sur un critère. Or, il y en a plus d'un (par exemple, la précision du résultat, le temps de calcul, l'occupation de mémoire nécessaire), et il arrive qu'on demande d'évaluer un algorithme selon plusieurs critères assez contradictoires.

En choisissant les algorithmes à examiner dans le présent livre, les auteurs sont essentiellement partis du critère de précision du résultat et de celui de vitesse de convergence du processus itératif. Même en restant dans ce cadre étroit, on ne peut cependant ordonner univoquement tous les algorithmes ni en indiquer le meilleur et le pire. Le fait est qu'on obtient les estimations de la vitesse de convergence pour des classes de problèmes, non pas pour des problèmes isolés, et un algorithme mauvais pour une vaste classe peut s'avérer efficace pour une autre, plus restreinte. Le calculateur doit donc posséder tout un arsenal d'algorithmes pour être en mesure de faire face à chaque problème proposé.

On a également à tenir compte de la façon dont on réalise une grande rapidité de convergence. Dans la pratique, même le calcul de dérivées premières d'une fonction s'avère souvent ardu et celui de dérivées secondes inextricable. Les auteurs ont donc insisté sur

des algorithmes qui exigent le calcul des seules dérivées premières ou des valeurs de la fonction.

Les auteurs se placent en dimension finie. Car, premièrement, en calcul automatique la solution d'un problème doit être approchée par la solution d'un autre en dimension finie, et, deuxièmement, la plupart des algorithmes se généralisent assez facilement à la minimisation de fonctionnelles sans subir des modifications tant soit peu essentielles. Les auteurs ont résolu par conséquent de se borner au cas en dimension finie. Le livre n'en est devenu que plus accessible au large public puisque la plupart des résultats ne nécessitent pour être compris que la connaissance des fondements de l'analyse mathématique et de l'algèbre linéaire.

Afin de ne pas alourdir l'exposé, les références dans le texte même sont très rares (elles sont en général réunies dans les brefs commentaires à la fin de chaque chapitre). La littérature relative aux questions traitées étant trop abondante, la bibliographie ne renferme pour l'essentiel que des articles et des monographies dont les auteurs se sont servis directement.

Le présent livre passe complètement sous silence les méthodes de résolution de la vaste classe importante de problèmes d'extrémum mal posés, méthodes élaborées par A. Tikhonov et son école. Les auteurs effleurent à peine la résolution de problèmes de commande optimale. Ces problèmes et les techniques de résolution correspondantes sont étudiés de divers points de vue dans la monographie de N. Moïsséev *Méthodes numériques en théorie des systèmes optimaux*.

Les algorithmes que le lecteur trouvera dans les pages ci-dessous sont de nature itérative. Cela signifie qu'on construit une suite finie ou infinie de points x_k, $k = 0, 1, \ldots$, dont on peut dire qu'elle converge dans un sens ou dans un autre vers la solution du problème de minimisation, les points successifs étant reliés par

$$x_{k+1} = x_k + \alpha_k p_k,$$

où p_k est le vecteur déplacement à partir du point x_k et α_k le pas dans la direction p_k. Aussi la description de tout algorithme de cet ouvrage consiste à définir le choix du vecteur p_k et la grandeur du pas α_k. Notons que si le choix de p_k définit la vitesse totale de convergence, celui de α_k influe sensiblement sur le volume de calcul à chaque itération. Si bien que les auteurs ont tâché de choisir α_k de façon qu'on puisse trouver la valeur cherchée du pas au bout d'un nombre fini d'opérations sans pour autant réduire la vitesse totale de convergence.

Voyons brièvement les estimations de la vitesse de convergence que les auteurs utilisent le plus souvent.

On dit que *la convergence* d'une suite $\{x_k\}$ vers un point x_* *est linéaire* ou que la suite *converge en progression géométrique* (de rapport q) si, à partir d'un certain k, on a l'inégalité $\| x_{k+1} - x_* \| \leqslant$

$\leqslant q \| x_k - x_* \|$, où $0 < q < 1$. Si l'on a l'inégalité $\| x_{k+1} - x_* \| \leqslant q_k \| x_k - x_* \|$, avec $q_k \to 0$ pour $k \to \infty$, on dit que *la convergence* de la suite $\{x_k\}$ *est superlinéaire* ou qu'elle *converge plus rapidement qu'en progression géométrique*. Si $q_k \leqslant C \| x_k - x_* \| \to 0$, alors $\| x_{k+1} - x_* \| \leqslant C \| x_k - x_* \|^2$. Cette estimation montre que *la convergence est quadratique*.

Le lecteur retrouvera plus bas ces estimations sous des formes équivalentes.

Quelques mots sur les notations.

Comme on l'a dit plus haut, les auteurs se placent dans un espace vectoriel n-dimensionnel qui sera noté E^n. On représentera un vecteur par une minuscule latine et ses coordonnées seront spécifiées par des indices supérieurs de sorte que x^i est la i-ième coordonnée du vecteur x. Les indices inférieurs désigneront les éléments d'une suite. On notera des matrices par des majuscules A, B, C, . . ., A^* désignant la transposée de A. On comprendra par x un vecteur-colonne de sorte que x^* sera un vecteur-ligne. Le produit scalaire de deux vecteurs sera représenté comme (x, y), i.e.

$$(x, y) = \sum_{i=1}^{n} x^i y^i.$$

Ceci étant, on considérera, sauf indication contraire, comme norme d'un vecteur sa norme euclidienne

$$\| x \| = \sqrt{(x, x)}.$$

Les auteurs tiennent à exprimer leurs remerciements les plus chaleureux à G. Lubarskaïa, L. Sobolenko, E. Bogouslavskaïa et à V. Panine pour leur aide inappréciable lors de la rédaction du présent ouvrage.

Les chapitres premier (à l'exception du § 5 et d'une partie du § 2) et III (sauf le § 9 et partiellement le § 3) sont dus à B. Pchénitchny.

Le chapitre II, les n°s 3–4 du § 2 et le § 5 du chapitre premier, les n°s 5, 6 du § 3 et le § 9 du chapitre III sont écrits par Y. Daniline.

B. Pchénitchny, Y. Daniline

CHAPITRE PREMIER

INTRODUCTION À LA THÉORIE DE LA PROGRAMMATION MATHÉMATIQUE

Les auteurs exposent certains résultats de la théorie des ensembles convexes et des conditions nécessaires d'extrémum, qui sont indispensables à la compréhension des chapitres suivants.

§ 1. Ensembles convexes

Dans ce paragraphe nous examinerons les propriétés fondamentales des ensembles convexes dans l'espace euclidien n-dimensionnel E^n.

1. Définition. Théorème de séparation.

DÉFINITION 1.1. *On dit qu'un ensemble X de points contenu dans E^n est convexe si, toutes les fois qu'il contient des points x_1, x_2 quelconques, il contient tous les points de la forme*

$$x = \lambda x_1 + (1 - \lambda) x_2, \ 0 \leqslant \lambda \leqslant 1.$$

En langage géométrique, cela signifie que si les extrémités d'un segment appartiennent à un ensemble convexe X, il en est de même du segment tout entier.

LEMME 1.1. *Les affirmations suivantes sont vraies:*

1. *L'intersection d'un nombre quelconque d'ensembles convexes est convexe.*

2. *Si $x_i \in X$, $i = 1, \ldots, m$, alors pour tous les λ_i, $i = 1, \ldots, m$, vérifiant la condition $\sum_{i=1}^{m} \lambda_i = 1$, $\lambda_i \geqslant 0$, $i = 1, \ldots, m$, le point $x = \sum_{i=1}^{m} \lambda_i x_i$ appartient à X.*

Le théorème suivant et ses conséquences sont cet instrument principal à l'aide duquel on obtient les résultats caractérisant telle ou telle propriété des ensembles convexes.

THÉORÈME 1.1. *Etant donné un ensemble X convexe, sa fermeture $\bar{X}$ et un point x_0 n'appartenant pas à $\bar{X}$, il existe un vecteur $a \in E^n$, $a \neq 0$, et un nombre $\varepsilon > 0$ tels que*

$$(a, x) \leqslant (a, x_0) - \varepsilon$$

pour tous les $x \in X$.

DÉMONSTRATION. $\bar{X}$ est par définition un ensemble fermé. Montrons qu'il est convexe. En effet, si $x \in \bar{X}$, il existe une suite x_k, $k = 1, \ldots$, telle que $x_k \in X$, $x_k \to x$. Soit maintenant $x, y \in \bar{X}$, $0 \leqslant \lambda \leqslant 1$. Montrons que $\lambda x + (1 - \lambda) y \in \bar{X}$. Etant donné la convexité de X, $x_k, y_k \in X$, $x_k \to x$, $y_k \to y$ entraîne

$$\lambda x_k + (1 - \lambda) y_k \in X,$$
$$\lambda x_k + (1 - \lambda) y_k \to \lambda x + (1 - \lambda) y.$$

Or, cela signifie justement que $\lambda x + (1 - \lambda) y \in \bar{X}$, i.e. $\bar{X}$ est convexe.

Prenons un point $y \in \bar{X}$ dont la distance à x_0 soit minimale, i.e.

$$\| x - x_0 \| \geqslant \| y - y_0 \|, \ x \in \bar{X}.$$

$\bar{X}$ étant convexe, on a, quels que soient $x \in X$ et $0 \leqslant \lambda \leqslant 1$

$$\lambda x + (1 - \lambda) y = y + \lambda (x - y) \in X.$$

Donc

$$\begin{aligned}
&\| \lambda x + (1 - \lambda) y - x_0 \|^2 = \| y - x_0 + \lambda (x - y) \|^2 = \\
&= (y - x_0 + \lambda (x - y), y - x_0 + \lambda (x - y)) = \\
&= (y - x_0, y - x_0) + 2\lambda (y - x_0, x - y) + \lambda^2 (x - y, x - y) = \\
&= \| y - x_0 \|^2 + 2\lambda (y - x_0, x - y) + \lambda^2 \| x - y \|^2 \geqslant \| y - x_0 \|^2.
\end{aligned}$$

La dernière inégalité est juste pour tous les λ variant de 0 à 1. Simplifions-la, nous obtenons

$$2 (y - x_0, x - y) + \lambda \| x - y \|^2 \geqslant 0,$$

d'où, lorsque $\lambda = 0$,

$$(y - x_0, x - y) \geqslant 0.$$

Posons $a = x_0 - y$. La dernière inégalité s'écrit alors comme $(a, x) \leqslant (a, y)$. Mais

$$(a, y) = (a, x_0) - (a, x_0 - y) = (a, x_0) - \| a \|^2.$$

Ainsi, en posant $\varepsilon = \| a \|^2$, on obtient finalement

$$(a, x) \leqslant (a, x_0) - \varepsilon.$$

Cette inégalité est vérifiée pour n'importe quels $x \in \overline{X}$. De plus, $\varepsilon > 0$ parce que $x_0 \notin \overline{X}$ et donc $y \neq x_0$. C'est pourquoi

$$\varepsilon = \| a \|^2 = \| x_0 - y \|^2 > 0.$$

Le théorème est démontré.

Remarque. En démontrant le théorème 1.1 nous avons prouvé par là même la convexité de la fermeture d'un ensemble convexe. A titre d'exercice simple on démontre que l'ensemble de points intérieurs d'un ensemble convexe est convexe.

Conséquence 1.1. *Soient X un ensemble convexe et x_0 un point frontière de X. Il existe alors un vecteur $a \neq 0$ tel que*

$$(a, x) \leqslant (a, x_0), \ x \in X.$$

Conséquence 1.2. *Si X et Y sont deux ensembles convexes sans point commun, il existe alors un vecteur $a \neq 0$ tel que*

$$(a, x) \leqslant (a, y), \ x \in X, \ y \in Y.$$

Conséquence 1.3. *Si X et Y sont deux ensembles convexes fermés sans point commun dont l'un est borné, il existe un vecteur $a \neq 0$ et un nombre $\varepsilon > 0$ tels que*

$$(a, x) \leqslant (a, y) - \varepsilon, \ x \in X, \ y \in Y.$$

2. Cônes convexes.

Definition 1.2. *Un ensemble K s'appelle cône convexe s'il est convexe et si, toutes les fois qu'il contient un point $x \in K$, il contient tous les points λx pour $\lambda > 0$.*

On voit aisément que si $x, y \in K$, alors $x + y \in K$. En effet, K étant convexe, le point $\frac{1}{2} x + \frac{1}{2} y$ appartient à K. Mais

$$x + y = 2\left(\frac{1}{2} x + \frac{1}{2} y\right),$$

d'où $x + y \in K$ par définition d'un cône. Les propriétés les plus importantes des cônes s'énoncent en termes de la dépendance entre un cône originel (primal) et son dual.

Definition 1.3. *Soit K un cône convexe. L'ensemble de tous les vecteurs $y \in E^n$ satisfaisant, quels que soient $x \in K$, à l'inégalité*

$$(x, y) \geqslant 0$$

s'appelle cône dual et se note K^*.

On vérifie immédiatement que K^* est également convexe.

Lemme 1.2. *K^* est un cône convexe fermé.*

LEMME 1.3. *Soit K un cône convexe. Alors $x_0 \in \overline{K}$ si et seulement si $(x_0, y) \geqslant 0$ pour tous les $y \in K^*$. Si K est fermé, alors*

$$(K^*)^* = K.$$

DÉMONSTRATION. Il est évident que, lorsque $x_0 \in \overline{K}$, on a $(x_0, y) \geqslant 0$ pour tous les $y \in K^*$. Procédons par l'absurde. Ainsi, soit $(x_0, y) \geqslant 0$ pour tous les $y \in K^*$, mais $x_0 \notin \overline{K}$. Vu que K est convexe fermé, il existe, en vertu du théorème 1.1, un vecteur a tel que

$$(a, x_0) \leqslant (a, x) - \varepsilon, \ x \in \overline{K}.$$

Or, un cône fermé K contient toujours le point 0. On a donc, en particulier,

$$(a, x_0) \leqslant -\varepsilon. \tag{1.1}$$

D'autre part,

$$(a, x) \geqslant 0, \ x \in \overline{K}. \tag{1.2}$$

En effet, si $(a, x_1) < 0$ pour un certain $x_1 \in K$, on a, du moment que $\lambda x_1 \in K$ pour $\lambda > 0$,

$$(a, x_0) \leqslant \lambda (a, x_1) - \varepsilon,$$

et la dernière inégalité est vérifiée pour tous les λ, chose impossible si $(a, x_1) < 0$. Ainsi, (1.2) est juste, et donc $a \in K^*$. Mais alors $(a, x_0) \geqslant 0$, ce qui contredit (1.1). La première affirmation du lemme vient d'être démontrée.

Démontrons la seconde affirmation. Si $x \in K$, alors, par définition, $(x, y) \geqslant 0$ quels que soient $y \in K^*$, donc $x \in (K^*)^*$, $K \subset (K^*)^*$. Inversement, par définition $x \in (K^*)^*$ si et seulement si $(x, y) \geqslant 0$ pour tous les $y \in K^*$. On a montré plus haut que dans ce cas $x \in K$, i.e. $(K^*)^* \subset K$. Ainsi, $(K^*)^* = K$, c.q.f.d.

Une classe importante de cônes de la théorie de la programmation linéaire est celle de cônes polyédriques.

DÉFINITION 1.4. *Un cône K est dit polyédrique s'il existe un ensemble fini de n-vecteurs a_i, $i = 1, \ldots, m$, tel que $x \in K$ entraîne le développement*

$$x = \sum_{i=1}^{m} \lambda_i a_i, \quad \lambda_i \geqslant 0, \ i = 1, \ldots, m, \tag{1.3}$$

et, inversement, (1.3) *implique $x \in K$.*

On voit donc qu'un cône polyédrique K est tout simplement un ensemble de points susceptibles de se mettre sous la forme (1.3). Pour un point $x \in K$ donné, la représentation (1.3). n'est en général pas unique.

LEMME 1.4. *Soit $x \in K$, K étant un cône polyédrique. Il existe alors un développement de x suivant les vecteurs a_i à coefficients non négatifs λ_i tel que le nombre d'indices i pour lesquels $\lambda_i > 0$ soit au plus égal à la dimension de l'espace n. Ceci étant, les vecteurs a_i correspondant à des λ_i non nuls sont linéairement indépendants.*

DEMONSTRATION. Soit $x \in K$, i.e. $x = \sum_{i=1}^{m} \lambda_i a_i$, et soit $\mathcal{J}$ l'ensemble d'indices i pour lesquels $\lambda_i > 0$. Admettons que le nombre d'éléments de $\mathcal{J}$ est au moins égal à n et que les vecteurs a_i, $i \in \mathcal{J}$, sont linéairement dépendants. Comme un espace n-dimensionnel contient n vecteurs linéairement indépendants au plus, on peut toujours indiquer des coefficients α_i non tous nuls tels que $\sum_{i \in \mathcal{J}} \alpha_i a_i = 0$. De plus, par définition de $\mathcal{J}$, $\lambda_i = 0$ si $i \notin \mathcal{J}$, de sorte que

$$x = \sum_{i \in \mathcal{J}} \lambda_i a_i, \quad \lambda_i > 0, \ i \in \mathcal{J}.$$

En retranchant la relation précédente et en multipliant par ε on obtient

$$x = \sum_{i \in \mathcal{J}} (\lambda_i - \varepsilon \alpha_i)\, a_i.$$

On peut estimer, sans restreindre la généralité, que $\alpha_i > 0$ pour certains $i \in \mathcal{J}$. Posons $\varepsilon_0 = \min_{i \in \mathcal{J},\, \alpha_i > 0} \frac{\lambda_i}{\alpha_i}$. En posant $\overline{\lambda_i} = \lambda_i - \varepsilon_0 \alpha_i$, il vient alors

$$x = \sum_{i \in \mathcal{J}} \overline{\lambda}_i a_i,$$

$\overline{\lambda}_i \geqslant 0$, l'égalité ayant lieu pour au moins un i.

Nous avons obtenu le développement à coefficients non négatifs de x suivant les vecteurs a_i, mais le nombre de coefficients strictement négatifs a diminué.

Continuons ce processus tant que le nombre de coefficients non nuls n'est pas $\leqslant n$ et les vecteurs a_i pour lesquels $\lambda_i > 0$ ne deviennent pas linéairement indépendants.

Du moment qu'il s'agit d'un entier, le processus ne peut durer indéfiniment, et, après un nombre fini de pas, on aboutit au développement satisfaisant aux hypothèses du lemme.

LEMME 1.5. *Un cône polyédrique est fermé.*

LEMME 1.6. *Soit K un cône donné par le système d'inégalités linéaires*

$$(a_i, x) \geqslant 0, \ i = 1, \ldots, m,$$

où $a_1 \in E^n$. Le cône dual K^* est alors polyédrique et se compose de points y de la forme

$$y = \sum_{i=1}^{m} \lambda_i a_i, \quad \lambda_i \geqslant 0, \quad i = 1, \ldots, m.$$

DÉMONSTRATION. Considérons le cône

$$\widetilde{K} = \{y\colon y = \sum_{i=1}^{m} \lambda_i a_i, \ \lambda_i \geqslant 0, \ i = 1, \ldots, m\}.$$

$\widetilde{K}^*$ est par définition un ensemble de points x pour lesquels $(x, y) \geqslant 0$, $y \in \widetilde{K}$, i.e. $(x, \sum_{i=1}^{m} \lambda_i a_i) \geqslant 0$ quels que soient $\lambda_i \geqslant 0$, auquel cas

$$(x, \sum_{i=1}^{m} \lambda_i a_i) = \sum_{i=1}^{m} \lambda_i (x, a_i) \geqslant 0.$$

Il est évident que la dernière inégalité n'est vérifiée pour n'importe quels $\lambda_i \geqslant 0$ que si $(a_i, x) \geqslant 0$, $i = 1, \ldots, m$, i.e. si $x \in K$. Donc $\widetilde{K}^* = K$. Etant donné que $\widetilde{K}$ est polyédrique, il est fermé et, en vertu du lemme 1.3, $(\widetilde{K}^*)^* = \widetilde{K}$. Ainsi, $K^* = \widetilde{K}$, c.q.f.d.

REMARQUE. Le lemme que nous venons de démontrer (lemme de Farkaš-Minkowski) constitue cet outil principal dont on se sert pour obtenir les conditions nécessaires d'extrémum.

3. Ensembles strictement et fortement convexes.

DÉFINITION 1.5. *Un ensemble $X \subset E^n$ est dit strictement convexe si, quels que soient $x_1, x_2 \in X$, $x_1 \neq x_2$, tous les points de la forme*

$$\lambda x_1 + (1 - \lambda) x_2, \quad 0 < \lambda < 1,$$

sont intérieurs à X.

DÉFINITION 1.6. *Un ensemble $X \subset E^n$ est dit fortement convexe s'il existe une constante $\gamma > 0$ telle que tout point*

$$\frac{x_1 + x_2}{2} + y \in X$$

si $x_1, x_2 \in X$ et $\|y\| \leqslant \gamma \|x_2 - x_1\|^2$.

On constate facilement qu'un ensemble fortement convexe l'est strictement (la réciproque n'étant pas juste).

§ 2. Fonctions convexes

Vu leurs propriétés importantes, les fonctions convexes constituent l'un des grands chapitres de la théorie de la programmation mathématique. C'est en termes de ces fonctions que s'énonce le problème de programmation convexe, le problème d'extrémum le plus étudié. Elles jouent également un rôle de premier plan dans le problème non linéaire général car on n'arrive à formuler des conditions nécessaires d'extrémum assez générales et exhaustives que si les dérivées directionnelles en un point donné des fonctions figurant dans le problème sont des fonctions convexes de la direction.

Les fonctions convexes que nous rencontrerons dans la suite sont pour la plupart définies dans tout l'espace si bien qu'étant donnée une fonction convexe, sa valeur est finie en chaque point $x \in E^n$. Du point de vue de la théorie générale il est des fois utile de considérer également des fonctions convexes qui peuvent être $+\infty$ en certains points. Nous ne les rencontrerons toutefois que fort rarement (en étudiant les problèmes duals de la programmation convexe) si bien que partout dans ce paragraphe on suppose, sauf indication du contraire, que la fonction convexe considérée est définie dans l'espace E^n tout entier et y prend des valeurs finies.

1. Définition. Propriétés fondamentales.

DEFINITION 2.1. *Une fonction $f(x)$ définie pour tous les $x \in E^n$ est dite convexe si*

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leqslant \lambda_1 f(x_1) + \lambda_2 f(x_2)$$

quels que soient x_1, x_2 et $\lambda_1, \lambda_2 \geqslant 0$, $\lambda_1 + \lambda_2 = 1$.

REMARQUE. Si $f(x) = +\infty$ pour certains x, la définition reste en vigueur.

LEMME 2.1. *Soient $f_1(x)$ et $f_2(x)$ deux fonctions convexes et c_1, c_2 deux nombres non négatifs. Alors*

$$f(x) = c_1 f_1(x) + c_2 f_2(x)$$

est également convexe

LEMME 2.2 *Soient $f_i(x)$, $i = 1, \ldots, m$, des fonctions convexes. Il en est alors également de $f(x) = \max\limits_{1 \leqslant i \leqslant m} f_i(x)$.*

LEMME 2 3 *Si $f(x)$ est une fonction convexe, alors*

$$f(\lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_m x_m) \leqslant \lambda_1 f(x_1) + \lambda_2 f(x_2) + \ldots + \lambda_m f(x_m)$$

pour tous les λ_i non négatifs vérifiant la condition

$$\lambda_1 + \ldots + \lambda_m = 1.$$

DEMONSTRATION. Lorsque $m = 2$, l'affirmation découle de la définition d'une fonction convexe. Supposons qu'elle est démontrée pour $m \leqslant k$. Montrons qu'elle est juste pour $m = k + 1$. Soit $\lambda_i \geqslant 0$, $i = 1, \ldots, k + 1$, $\lambda_1 + \ldots + \lambda_{k+1} = 1$. On peut dire évidemment que tous les λ_i sont strictement supérieurs à 0 sinon l'inégalité serait vérifiée par hypothèse.

Ainsi, $\lambda_{k+1} > 0$ et $1 - \lambda_{k+1} = \lambda_1 + \ldots + \lambda_k > 0$.

On a par définition d'une fonction convexe

$$f(\lambda_1 x_1 + \ldots + \lambda_k x_k + \lambda_{k+1} x_{k+1}) \leqslant$$
$$\leqslant (1 - \lambda_{k+1}) f\left(\frac{\lambda_1}{1-\lambda_{k+1}} x_1 + \ldots + \frac{\lambda_k}{1-\lambda_{k+1}} x_k\right) + \lambda_{k+1} f(x_{k+1}). \quad (2.1)$$

Mais d'après l'hypothèse de récurrence

$$f\left(\frac{\lambda_1}{1-\lambda_{k+1}} x_1 + \ldots + \frac{\lambda_k}{1-\lambda_{k+1}} x_k\right) \leqslant$$
$$\leqslant \frac{\lambda_1}{1-\lambda_{k+1}} f(x_1) + \ldots + \frac{\lambda_k}{1-\lambda_{k+1}} f(x_k), \quad (2.2)$$

car

$$\frac{\lambda_1}{1-\lambda_{k+1}} + \ldots + \frac{\lambda_k}{1-\lambda_{k+1}} = 1.$$

La comparaison des inégalités (2.1) et (2.2) donne le résultat cherché. Le lemme est démontré.

LEMME 2.4. *Une fonction $f(x)$ est convexe si et seulement si il en est également de la fonction d'argument t unidimensionnel quels que soient x et $p \in E^n$*

$$\varphi_{x,\,p}(t) \equiv f(x + tp). \quad (2.3)$$

2. Propriétés de dérivabilité. Soit $f(x)$ une fonction convexe continuement dérivable à gradient $f'(x)$.

LEMME 2.5. *Les affirmations suivantes sont équivalentes:*

1) *$f(x)$ est une fonction convexe.*

2) *$f(x_2) - f(x_1) \geqslant (f'(x_1), x_2 - x_1)$ quels que soient $x_1, x_2 \in E^n$.*

3) *$(f'(x + \lambda p), p)$ est une fonction non décroissante de λ.*

Si $f(x)$ est deux fois continûment dérivable, alors

4) *la matrice des dérivées secondes $f''(x)$ est définie positive, i.e. $(f''(x)\,p, p) \geqslant 0$ pour tous les $x, p \in E^n$.*

DEMONSTRATION. Commençons par remarquer que si

$$\varphi_{x,\,p}(\lambda) = f(x + \lambda p),$$

alors $\varphi_{x,\,p}(\lambda)$ est convexe (on l'a montré plus haut) et

$$\varphi'_{x,\,p}(\lambda) = (f'(x + \lambda p), p), \quad \varphi''_{x,\,p}(\lambda) = (p, f''(x + \lambda p)\,p). \quad (2.4)$$

Montrons que l'affirmation 1 entraîne 2. En effet, vu que

$$f((1-\lambda)x_1+\lambda x_2)\leqslant(1-\lambda)f(x_1)+\lambda f(x_2),\ 0<\lambda<1,$$

on a

$$\frac{f(x_1+\lambda(x_2-x_1))-f(x_1)}{\lambda}\leqslant f(x_2)-f(x_1).$$

Le passage à la limite pour $\lambda\to 0$ donne

$$(f'(x_1),\ x_2-x_1)\leqslant f(x_2)-f(x_1).$$

Donc l'affirmation 1 entraîne 2, ou, plus brièvement, $1\to 2$.

Montrons que $2\to 3$. En vertu de 2, on a pour $\varphi_{x,p}(\lambda)$

$$\varphi'_{x,p}(\lambda_1)(\lambda_2-\lambda_1)\leqslant\varphi_{x,p}(\lambda_2)-\varphi_{x,p}(\lambda_1),$$
$$\varphi'_{x,p}(\lambda_2)(\lambda_1-\lambda_2)\leqslant\varphi_{x,p}(\lambda_1)-\varphi_{x,p}(\lambda_2).$$

Ces deux inégalités donnent quand $\lambda_2>\lambda_1$

$$\varphi'_{x,p}(\lambda_1)\leqslant\frac{\varphi_{x,p}(\lambda_2)-\varphi_{x,p}(\lambda_1)}{\lambda_2-\lambda_1}\leqslant\varphi'_{x,p}(\lambda_2),$$

i.e. $(f'(x+\lambda_1 p),\ p)\leqslant(f'(x+\lambda_2 p),\ p)$, c.q.f.d.

$3\to 1$. Soit $(f'(x+\lambda p),\ p)$ une fonction non décroissante de λ. Alors $\varphi'_{x,p}(\lambda_1)\leqslant\varphi'_{x,p}(\lambda_2)$ pour $\lambda_2\geqslant\lambda_1$. Si $0<\mu<1$, alors

$$0\leqslant\mu(\lambda_2-\lambda_1)\int_0^1[\varphi'_{x,p}(\lambda_1+\tau(\lambda_2-\lambda_1))-\varphi'_{x,p}(\lambda_1+\tau\mu(\lambda_2-\lambda_1))]\,d\tau=$$
$$=(1-\mu)\varphi_{x,p}(\lambda_1)+\mu\varphi(\lambda_2)-\varphi_{x,p}((1-\mu)\lambda_1+\mu\lambda_2),$$

i.e. $\varphi_{x,p}(\lambda)$ est une fonction convexe de λ. Il découle alors du lemme 2.4 que $f(x)$ est convexe.

$3\to 4$. Comme $\varphi'_{x,p}(\lambda)=(f'(x+\lambda p),\ p)$ est une fonction non décroissante, on a $\varphi''_{x,p}(\lambda)\geqslant 0$, i.e.

$$(p,\ f''(x+\lambda p)\,p)\geqslant 0, \tag{2.5}$$

d'où la définie positivité de la matrice $f''(x)$.

$4\to 3$. Inversement, si (2.5) est vérifié, alors $\varphi''_{x,p}(\lambda)$ est non négative et donc $\varphi'_{x,p}(\lambda)=(f'(x+\lambda p),\ p)$ est non décroissante.

Nous avons montré que $1\to 2\to 3\to 1$, $4\to 3$ et $3\to 4$, ce qui signifie l'équivalence des quatre assertions du lemme.

Conséquence 2.1. *Une fonction quadratique*

$$f(x)=\frac{1}{2}(x,\ Ax)+(b,\ x)$$

est convexe si et seulement si la matrice A est définie positive.

En effet, $f(x)$ est deux fois continûment dérivable et $f''(x)=A$. La conséquence découle donc de suite de l'affirmation 4 du lemme 2.5.

Le lemme démontré fournit plusieurs critères de convexité d'une fonction.

DÉFINITION 2.2. *Soit $f(x)$ une fonction convexe définie en un point x_0 et y prenant une valeur finie. Un vecteur g s'appelle sous-gradient ou vecteur d'appui en x_0 à $f(x)$ si on a, quel que soit x,*

$$f(x) - f(x_0) \geqslant (g, x - x_0). \tag{2.6}$$

On montre que si $f(x)$ est continue en x_0, il existe en ce point des sous-gradients et l'ensemble des sous-gradients est convexe, fermé et borné. Le lemme 2.5 (affirmation 2)) implique que $f'(x_0)$ est un sous-gradient de $f(x)$ en x_0 si cette dernière est une fonction dérivable. La notion de sous-gradient généralise donc celle de gradient.

La définition aidant, on voit sans peine que si g_1 et g_2 sont des sous-gradients des fonctions convexes $f_1(x)$ et $f_2(x)$ au point x_0, alors $c_1g_1 + c_2g_2$ est un sous-gradient de la fonction $c_1f_1(x) + c_2f_2(x)$, $c_1, c_2 \geqslant 0$. Connaissant les sous-gradients de fonctions convexes on calcule donc sans difficulté le sous-gradient de leur combinaison linéaire.

Soit maintenant $f(x) = \max\limits_{i=1,\ldots,m} f_i(x)$, où $f_i(x)$ est convexe, et soient g_i les sous-gradients de $f_i(x)$ en x_0. Le vecteur g

$$g = \sum_{i=1}^{m} \lambda_i g_i,$$

$$\sum_{i=1}^{m} \lambda_i = 1,\ \lambda_i \geqslant 0,\ i = 1, \ldots, m,\ \lambda_i = 0, \text{ si } f_i(x_0) < f(x_0),$$

est un sous-gradient de la fonction $f(x)$.

3. Fonctions strictement et fortement convexes. Les fonctions remplissant une condition renforcée de convexité jouent un rôle ort important en programmation mathématique.

DÉFINITION 2.3. *Une fonction $f(x)$ est dite strictement convexe si*

$$f((1-\lambda)x + \lambda y) < (1-\lambda) f(x) + \lambda f(y),\ 0 < \lambda < 1,\quad x \neq y.$$

Pour une fonction strictement convexe suffisamment régulière on a les assertions analogues à celles du lemme 2.5.

LEMME 2.6. *Les affirmations suivantes sont équivalentes:*

1) *$f(x)$ est une fonction strictement convexe.*

2) *$f(x_2) - f(x_1) > (f'(x_1), x_2 - x_1)$ quels que soient $x_1, x_2 \in E^n$, $x_1 \neq x_2$.*

3) *$(f'(x + \lambda p), p)$ est une fonction strictement croissante de λ.*

DÉFINITION 2.4. *Une fonction $f(x)$ est dite fortement convexe si, quels que soient $x_1, x_2 \in E^n$,*

$$f\left(\frac{x_1+x_2}{2}\right) \leqslant \frac{1}{2}[f(x_1)+f(x_2)] - \gamma \| x_2 - x_1 \|^2, \tag{2.7}$$

où $\gamma > 0$ est une constante aussi petite qu'on le veut.

Une fonction fortement convexe est, on le constate aisément, strictement convexe, mais la réciproque n'est en général pas vraie.

Dans la suite nous aborderons des fonctions fortement convexes deux fois continûment dérivables.

LEMME 2.7. *Si $f(x)$ est une fonction deux fois continûment dérivable, la condition de convexité forte* (2.7) *équivaut à*

$$(f''(x)\,p,\ p) \geqslant m \| p \|^2,\ m > 0, \tag{2.8}$$

quels que soient x et $p \in E^n$.

La dernière inégalité signifie que la matrice $f''(x)$ est *fortement positive.*

CONSÉQUENCE 2.2. *Une fonction quadratique strictement convexe $f(x) = \frac{1}{2}(Ax,\ x) + (b, x)$ définie dans E^n est fortement convexe et inversement.*

DÉMONSTRATION. Il ne faut prouver que la première affirmation.

Moyennant la condition 2) du lemme 2.6 on se convainc que pour tout $x \neq 0$

$$(Ax,\ x) > 0. \tag{2.9}$$

Par ailleurs,

$$(Ax,\ x) \geqslant \lambda\,(x,\ x) = \lambda \| x \|^2, \tag{2.10}$$

avec λ la plus petite valeur propre de la matrice des dérivées secondes A. La comparaison de (2.9) et (2.10) donne $\lambda > 0$. Ceci étant, la condition (2.10) signifie que $f(x)$ est fortement convexe.

Soit x_0 un point arbitraire de E^n. Considérons l'ensemble

$$Y = \{ x:\ f(x) \leqslant f(x_0) \}.$$

LEMME 2.8. *Si $f(x)$ est une fonction fortement convexe deux fois continûment dérivable, alors Y est un ensemble fortement convexe, fermé et borné.*

DÉMONSTRATION. La fermeture de Y découle de la continuité de $f(x)$.

Démontrons que Y est borné. Selon la formule de Taylor

$$f(x) = f(x_0) + (f'(x_0),\ x - x_0) + \frac{1}{2}(f''(\xi)(x - x_0),\ x - x_0),$$

où $\xi = x_0 + \theta(x - x_0)$, $\theta \in [0, 1]$. Compte tenu de (2.8)

$$f(x_0) \geqslant f(x) \geqslant f(x_0) + (f'(x_0), x - x_0) + \frac{m}{2} \| x - x_0 \|^2.$$

Il en résulte

$$\frac{m}{2} \| x - x_0 \|^2 + (f'(x_0), x - x_0) < 0,$$

i.e.

$$\frac{m}{2} \| x - x_0 \|^2 \leqslant | (f'(x_0), x - x_0 | \leqslant \| f'(x_0) \| \| x - x_0 \|,$$

ou

$$\| x - x_0 \| \leqslant 2 \frac{\| f'(x_0) \|}{m},$$

ce qui prouve le caractère borné de Y.

Etablissons enfin la convexité forte de Y. Soit $x_1, x_2 \in Y$. Utilisons la formule de Lagrange et la condition (2.7), il vient

$$f\left(\frac{x_1 + x_2}{2} + y\right) = f\left(\frac{x_1 + x_2}{2}\right) + (f'(\xi), y) \leqslant$$

$$\leqslant \frac{1}{2} [f(x_1) + f(x_2)] - \gamma \| x_1 - x_2 \|^2 + M \| y \|. \qquad (2.11)$$

Ici $\xi = \frac{x_1 + x_2}{2} + \theta y$, $\theta \in [0, 1]$, M étant la valeur maximale de $f'(x)$ sur Y.

Supposons que $f(x_1) \geqslant f(x_2)$. Alors $\frac{1}{2}[f(x_1) + f(x_2)] \leqslant f(x_1)$. Si $\| y \| \leqslant \frac{\gamma}{M} \| x_2 - x_1 \|^2$, il résulte de (2.11) que $f\left(\frac{x_1 + x_2}{2} + y\right) \leqslant f(x_1)$, i.e. $\frac{x_1 + x_2}{2} + y \in Y$. Par définition cela signifie que Y est fortement convexe.

Le lemme est démontré.

Remarque. La fermeture et la convexité forte de Y se conservent dans le cas où $f(x)$ est une fonction fortement convexe continue ou dérivable. On démontre alors la dernière propriété de Y moyennant le fait qu'une fonction fortement convexe continue vérifie sur tout ensemble borné la condition de Lipschitz (voir N. Bourbaki [43]).

Lemme 2.9. *Si la matrice $f''(x)$ vérifie la condition* (2.8), *il existe l'inverse $f''^{-1}(x)$ et*

$$(f''^{-1}(x) p, p) \leqslant \frac{1}{m} \| p \|^2.$$

Si la matrice $f''(x)$ est de plus bornée, i.e.

$$(f''(x) p, p) \leqslant M \| p \|^2, \qquad (2.12)$$

alors

$$(f''^{-1}(x)\,p,\,p) \geqslant \frac{m}{M^2}\,\|p\|^2.$$

4. Fonctions concaves.

DÉFINITION. *Si, quels que soient* x_1, $x_2 \in E^n$ *et* $0 \leqslant \lambda \leqslant 1$, *on a l'inégalité*

$$f(\lambda x_1 + (1-\lambda)\,x_2) \geqslant \lambda f(x_1) + (1-\lambda)\,f(x_2),$$

la fonction $f(x)$ *est dite concave.*

Il en découle que $f(x)$ est concave si et seulement si $-f(x)$ est convexe. Toutes les propriétés des fonctions concaves s'obtiennent donc sans difficulté des propriétés correspondantes des fonctions convexes.

Par analogie avec ces dernières on définit des fonctions strictement et fortement concaves et on étudie leurs propriétés.

§ 3. Programmation convexe

La programmation convexe a pour tâche de rechercher le minimum d'une fonction convexe dans un domaine convexe. La programmation convexe est le chapitre le mieux élaboré de la programmation mathématique.

1. Position du problème. Propriétés fondamentales. Etant donné une fonction continue convexe $f(x)$, $x \in E^n$, définie pour tous les $x \in E^n$ et un ensemble convexe X, on demande le minimum de $f(x)$ sur X, i.e. un point x_* tel que

$$f(x_*) \leqslant f(x),\ x \in X.$$

LEMME 3.1. *Une fonction continue convexe* $f(x)$ *atteint son minimum sur un convexe compact* X.

DÉMONSTRATION. L'affirmation du lemme n'est autre qu'un cas particulier du théorème connu de Weierstrass aux termes duquel une fonction continue atteint son minimum sur un compact.

LEMME 3.2. *Etant donné un ensemble* X *fermé et une fonction* $f(x)$ *fortement convexe deux fois continûment dérivable,* $f(x)$ *atteint son minimum sur* X.

DÉMONSTRATION. Soit $x_0 \in X$. Prenons l'ensemble

$$Y = \{x : f(x) \leqslant f(x_0)\}.$$

Il est borné et fermé en vertu du lemme 2.8.

Considérons l'intersection $X \cap Y$. Il est évident que si x_* réalise le minimum de $f(x)$ sur $X \cap Y$, c'est également un point de minimum

de $f(x)$ sur X. Mais l'ensemble $X \cap Y$ est borné et fermé en tant qu'intersection de deux ensembles fermés dont l'un est borné. Aussi $f(x)$ atteint son minimum sur $X \cap Y$ et, partant, sur X tout entier.

Toute fonction convexe et strictement convexe n'atteint pas son minimum.

LEMME 3.3. *Un ensemble $X_* \subset X$ des points réalisant le minimum d'une fonction convexe $f(x)$ sur X est convexe.*

LEMME 3.4. *Une fonction strictement convexe atteint son minimum sur un ensemble convexe X en un point unique.*

DÉMONSTRATION. Soient x_1 et x_2 deux points de minimum de $f(x)$ sur X. Alors

$$f\left(\frac{1}{2}x_1+\frac{1}{2}x_2\right)<\frac{1}{2}f(x_1)+\frac{1}{2}f(x_2)=f(x_1),$$
$$\frac{1}{2}x_1+\frac{1}{2}x_2\in X,$$

ce qui contredit le fait que x_1 réalise le minimum de $f(x)$.

2. Conditions nécessaires de minimum. Soient $f(x)$ une fonction convexe continûment dérivable et X un ensemble convexe. Une question se pose : si x_* est un point de minimum de $f(x)$ sur X, à quelles conditions ce point est-il astreint?

DÉFINITION 3.1. *Soit $x_0 \in X$. Notons $K(x_0)$ l'ensemble de vecteurs p tels que $p \in K(x_0)$ si et seulement si il existe un $\alpha > 0$ vérifiant $x_0 + \alpha p \in X$.*

L'ensemble $K(x_0)$ s'appelle cône des directions admissibles pour X en x_0.

LEMME 3.5. *$K(x_0)$ est un cône convexe. Si $p \in K(x_0)$ et $x_0 + \alpha_0 p \in X$, alors $x_0 + \alpha p \in X$ quels que soient $0 \leqslant \alpha \leqslant \alpha_0$.*

THÉORÈME 3.1. *Soit x_* un point en lequel une fonction convexe continûment dérivable $f(x)$ atteint son minimum sur un ensemble convexe X. On a*

$$f'(x_*) \in K^*(x_*). \tag{3.1}$$

Inversement, si (3.1) est rempli, x_* réalise le minimum de $f(x)$ sur X.

DÉMONSTRATION. Supposons (3.1) vérifié en x_*. Alors $(f'(x_*), p) \geqslant 0$, $p \in K(x_*)$. Si, sous cette condition, $x \in X$, on a $p = x - x_* \in K(x_*)$ parce que $x_* + (x - x_*) = x \in X$. Donc

$$(f'(x_*),\ x - x_*) \geqslant 0,\ x \in X.$$

Or, pour une fonction convexe selon le lemme 2.5,

$$f(x) - f(x_*) \geqslant (f'(x_*),\ x - x_*).$$

Aussi

$$f(x) - f(x_*) \geqslant 0,\ x \in X.$$

Il en résulte que x_* est un point de minimum de $f(x)$ sur X.

Démontrons que (3.1) est une condition nécessaire. Soit x_* un point de minimum. Quels que soient $x \in X$ et λ, $0 < \lambda \leqslant 1$, on a alors

$$f((1-\lambda)x_* + \lambda x) = f(x_* + \lambda(x - x_*)) \geqslant f(x_*),$$

ou

$$\frac{f(x_* + \lambda(x - x_*)) - f(x_*)}{\lambda} \geqslant 0.$$

En passant à la limite $\lambda \to 0$, on obtient

$$(f'(x_*),\ x - x_*) \geqslant 0,\ \ x \in X. \tag{3.2}$$

Soit maintenant $p \in K(x_*)$. Alors $x_* + \alpha p = x \in X$, $\alpha > 0$, ou

$$p = \frac{1}{\alpha}(x - x_*).$$

Alors

$$(f'(x_*),\ p) = \frac{1}{\alpha}(f'(x_*),\ x - x_*) \geqslant 0, \tag{3.3}$$

inégalité qui tient compte de (3.2) et de $\alpha > 0$.

Ainsi, l'inégalité (3.3) est vraie pour tout $p \in K(x_*)$, d'où $f'(x_*) \in K^*(x_*)$.

Consequence 3.1. *Dans les conditions du théorème, le point x_* réalise le minimum de $f(x)$ sur X si et seulement si l'on a l'inégalité*

$$(f'(x_*),\ x - x_*) \geqslant 0,\ x \in X.$$

En effet, nous venons de démontrer l'équivalence de (3.2) et (3.1).

Appliquons le théorème 3.1 au cas où X est défini par un système d'inégalités linéaires.

Soient donnés les vecteurs $a_i \in E^n$, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, $\mathcal{J}^-$ et $\mathcal{J}^0$ étant deux ensembles finis d'indices, et les nombres b_i correspondants. Soit X un domaine défini par le système suivant d'égalités et d'inégalités

$$(a_i,\ x) - b_i \leqslant 0,\ i \in \mathcal{J}^-,\ (a_i,\ x) - b_i = 0,\ \ i \in \mathcal{J}^0. \tag{3.4}$$

Décrivons le cône $K(x_0)$ en un point arbitraire $x_0 \in X$. Posons

$$\mathcal{J}^-(x_0) = \{i\colon (a_i,\ x_0) - b_i = 0,\ i \in \mathcal{J}^-\}.$$

Par définition $p \in K(x_0)$ si $x_0 + \alpha p \in X$ pour des α suffisamment petits. Or, on voit aisément que $x_0 + \alpha p \in X$, i.e. le point $x_0 + \alpha p$ vérifie (3.4) pour α petit si et seulement si

$$(a_i, p) \leqslant 0, \ i \in \mathcal{J}^-(x_0), \ (a_i, p) = 0, \ i \in \mathcal{J}^0. \tag{3.5}$$

Ainsi, le cône $K(x_0)$ est décrit par le système (3.5) qui se met sous forme équivalente

$$(-a_i, p) \geqslant 0, \ i \in \mathcal{J}^-(x_0), \ (a_i, p) \geqslant 0, \ i \in \mathcal{J}^0, \ (-a_i, p) \geqslant 0, \ i \in \mathcal{J}^0.$$

En vertu du lemme 1.6, le vecteur $y \in K^*(x_0)$ se récrit

$$y = \sum_{i \in \mathcal{J}^-(x_0)} -u^i a_i + \sum_{i \in \mathcal{J}^0} -u^{+i} a_i + \sum_{i \in \mathcal{J}^0} u^{-i} a_i,$$

avec u^i, u^{+i}, u^{-i} des nombres non négatifs. Notons $u^i = u^{+i} - u^{-i}$, $i \in \mathcal{J}^0$, il vient

$$y = -\sum_{i \in \mathcal{J}^-(x_0)} u^i a_i - \sum_{i \in \mathcal{J}^0} u^i a_i, \ u^i \geqslant 0, \ i \in \mathcal{J}^-(x_0). \tag{3.6}$$

THÉORÈME 3.2. *Soient $f(x)$ une fonction convexe dérivable et X un ensemble décrit par le système* (3.4). *Pour que le point x_* réalise le minimum de $f(x)$ sur X, il faut et il suffit qu'il existe des nombres u^i, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, tels que*

$$f'(x_*) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i = 0, \ u^i \geqslant 0, \ i \in \mathcal{J}^-, \ u^i = 0,$$

si $(a_i, x_*) - b_i < 0, i \in \mathcal{J}^-$.

DÉMONSTRATION. Le résultat est immédiat à condition d'utiliser le théorème 3.1 et la représentation (3.6) des éléments de $K^*(x_0)$ et si on pose $u^i = 0$ pour $i \notin \mathcal{J}^-(x_*)$.

CONSÉQUENCE 3.2. *Pour que le point x_* réalise le minimum d'une fonction convexe dérivable dans l'espace tout entier, il faut et il suffit qu'on ait l'égalité*

$$f'(x_*) = 0.$$

CONSÉQUENCE 3.3. *Pour que le point x_* réalise le minimum d'une fonction convexe dérivable sur l'ensemble*

$$x^j \geqslant 0, \ j \in \mathcal{Y},$$

avec $\mathcal{Y}$ un sous-ensemble des $j = 1, 2, \ldots, n$, il faut et il suffit qu'on ait les relations

$$\frac{\partial f(x_*)}{\partial x^j} \geqslant 0 \quad \text{si } x^j = 0, \ j \in \mathcal{Y},$$

$$\frac{\partial f(x_*)}{\partial x^j} = 0 \quad \text{si } x^j \neq 0 \text{ ou } j \notin \mathcal{Y}.$$

3. Théorème de Kuhn et Tucker. Les conditions nécessaires et suffisantes de minimum que nous venons d'examiner s'appuient sur une description abstraite d'un ensemble admissible X sur lequel on minimise $f(x)$. Dans une vaste classe de problèmes l'ensemble X est défini comme un système d'inégalités et d'égalités. Le n° 3 donne justement les conditions nécessaires de minimum pour ce cas plus concret.

Soient donnés donc des fonctions convexes $f_i(x)$, $i = 0, 1, \ldots, m$, et un ensemble convexe X. On demande de minimiser $f_0(x)$ avec les contraintes

$$f_i(x) \leqslant 0, \quad i = 1, \ldots, m, \quad x \in X. \tag{3.7}$$

THEOREME 3.3 (de Kuhn et Tucker). *Etant donné x_*, point de minimum de $f_0(x)$ sous les contraintes* (3.7), *s'il existe un point $x_1 \in X$ tel que*

$$f_i(x_1) < 0, \quad i = 1, \ldots, m,$$

il existe alors des nombres $u^i \geqslant 0$, $i = 1, \ldots, m$, tels que

$$\begin{gathered} f_0(x_*) + \sum_{i=1}^{m} u^i f_i(x_*) \leqslant f_0(x) + \sum_{i=1}^{m} u^i f_i(x), \quad x \in X, \\ u^i f_i(x_*) = 0, \quad i = 1, \ldots, m. \end{gathered} \tag{3.8}$$

Les conditions citées sont également suffisantes.

DEFINITION 3.2. *Les nombres u^i figurant dans le théorème s'appellent multiplicateurs de Lagrange*

4. Problème dual. Reprenons le problème de minimisation de la fonction convexe $f_0(x)$ avec les contraintes (3.7). Supposons $u^i \geqslant 0$, $i = 1, \ldots, m$, fixes. Calculons

$$\varphi(u) = \inf_{x \in X} \left[f_0(x) + \sum_{i=1}^{m} u^i f_i(x) \right]. \tag{3.9}$$

Ainsi, la fonction $\varphi(u)$ (qui peut, à vrai dire, prendre les valeurs $-\infty$) est définie pour $u \geqslant 0$. Nous laissons au lecteur le soin de s'assurer de la concavité de $\varphi(u)$.

TEOREME 3.4. *Etant donné $u \geqslant 0$ et x vérifiant les contraintes* (3.7), *on a*

$$\varphi(u) \leqslant f_0(x).$$

Si on a les conditions du théorème 3.3, *alors*

$$\max_{u \geqslant 0} \varphi(u) = \min_{x \in D} f_0(x),$$

où D est l'ensemble de points x vérifiant (3.7).

DEMONSTRATION. On a pour $u \geqslant 0$, $x \in D$

$$\varphi(u) \leqslant f_0(x) + \sum_{i=1}^{m} u^i f_i(x) \leqslant f_0(x).$$

Supposons maintenant que nous sommes dans les conditions du théorème 3.3. Il existe alors un vecteur $u_0 \geqslant 0$ tel qu'il vérifie les relations (3.8). Mais ces dernières entraînent

$$\varphi(u_0) = f_0(x_*) + \sum_{i=1}^{m} u_0^i f_i(x_*) = f_0(x_*).$$

Comme $\varphi(u) \leqslant f_0(x_*)$, il en résulte que le vecteur u_0 réalise le maximum de $\varphi(u)$ dans le domaine $u \geqslant 0$ et

$$\max_{u \geqslant 0} \varphi(u) = \varphi(u_0) = f_0(x_*) = \min_{x \in D} f_0(x),$$

c.q.f.d.

Le problème de maximisation de $\varphi(u)$ sous la contrainte $u \geqslant 0$ s'appelle problème dual de programmation convexe et u vecteur-variable dual.

Le théorème 3.4 s'interprète maintenant comme suit : dans les conditions du théorème de Kuhn et Tucker la valeur maximale de la fonction économique du problème dual coïncide avec la valeur minimale de celle du problème primal, les multiplicateurs de Lagrange de ce dernier étant solution du problème dual.

Le problème de programmation convexe se pose souvent comme suit : minimiser $f_0(x)$ sous les contraintes

$$f_i(x) \leqslant 0, \quad i \in \mathcal{J}^-, \quad f_i(x) = 0, \quad i \in \mathcal{J}^0, \quad x \in X. \tag{3.10}$$

Ici $\mathcal{J}^-$ et $\mathcal{J}^0$ sont deux ensembles finis d'indices, $f_0(x)$, $f_i(x)$, $i \in \mathcal{J}^-$, des fonctions de x convexes, $f_i(x)$, $i \in \mathcal{J}^0$, des fonctions linéaires et X un ensemble convexe.

Le dual de ce problème consiste à maximiser $\varphi(u)$ dans les contraintes $u^i \geqslant 0$, $i \in \mathcal{J}^-$, u possédant les composantes u^i, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, et

$$\varphi(u) = \inf_{x \in X} \Big[f_0(x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i(x)\Big]. \tag{3.11}$$

Ainsi, le nombre de variables duales est égal à celui de contraintes (3.10) et la variable u^i correspondant à la i-ième contrainte prend des valeurs non négatives s'il s'agit d'une contrainte inégalité et des valeurs de signe quelconque dans le cas d'une contrainte égalité.

5. Problème de programmation linéaire. Le problème de programmation linéaire consiste à minimiser une fonction $f_0(x) = (a_0, x)$

avec les contraintes (3.4)

$$(a_i, x) - b_i \leqslant 0, \ i \in \mathcal{J}^-, \qquad (a_i, x) - b_i = 0, \ i \in \mathcal{J}^0.$$

Ce problème coïncide avec (3.10) si

$$f_i(x) = (a_i, x) - b_i, \ X = E^n.$$

LEMME 3.6. *Si les contraintes* (3.4) *sont compatibles, ou bien le problème de programmation linéaire a x_* pour solution, ou bien la borne inférieure de $f_0(x) = (a_0, x)$ sous les contraintes* (3.4) *vaut* $-\infty$.

La démonstration de ce lemme figure dans les cours de programmation linéaire.

Les conditions nécessaires caractérisant x_*, solution du problème de programmation linéaire, sont fournies par le théorème 3.2 énoncé de façon adéquate puisque $f_0'(x) = a_0$.

THEOREME 3.5. *Pour que le point x_* soit solution du problème de programmation linéaire, il faut et il suffit qu'il existe des nombres u^i, $i \in \mathcal{J}^- \cup \mathcal{J}^0$ tels que*

$$a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i = 0, \ u^i \geqslant 0, \ i \in \mathcal{J}^-, \ u^i = 0, \tag{3.12}$$

si $(a_i, x) - b_i < 0$, $i \in \mathcal{J}^-$.

Construisons le dual du problème de programmation linéaire. On a par définition

$$\varphi(u) = \inf_{x \in E^n} [f_0(x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i(x)] =$$

$$= \inf_{x \in E^n} [(a_0, x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i ((a_i, x) - b_i)] =$$

$$= \inf_{x \in E^n} [((a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i), x) -$$

$$- \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i] = \begin{cases} -\sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i & \text{si } a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i = 0, \\ -\infty & \text{si } a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i \neq 0. \end{cases}$$

Ainsi, le problème dual, i.e. maximiser $\varphi(u)$ dans les conditions $u^i \geqslant 0$, $i \in \mathcal{J}^-$, équivaut à maximiser

$$-\sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i \tag{3.13}$$

sous les contraintes

$$a_0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i = 0, \ u^i \geqslant 0, \ i \in \mathcal{J}^-. \tag{3.14}$$

THÉORÈME 3.6. *Si le problème primal de programmation linéaire a une solution, les multiplicateurs de Lagrange sont solution du problème dual et la valeur minimale de la fonction économique du primal est égale à la valeur maximale de celle du dual.*

A part les contraintes (3.4) le problème de programmation linéaire contient souvent des contraintes de la forme

$$x^j \geqslant 0, \quad j \in \mathcal{Y}, \tag{3.4'}$$

où $\mathcal{Y}$ est un sous-ensemble des indices 1, 2, . . ., n . Moyennant le théorème précédent le lecteur démontrera sans peine le

THÉORÈME 3.7. *Si le problème de programmation linéaire avec les contraintes* (3.4), (3.4') *possède une solution, les multiplicateurs de Lagrange correspondant aux contraintes* (3.4) *sont solution du problème dual qui consiste à maximiser*

$$-\sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i$$

sous les contraintes

$$a_0^j + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i^j \geqslant 0, \quad j \in \mathcal{Y},$$

$$a_0^j + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i^j = 0, \quad j \notin \mathcal{Y}, \quad u^i \geqslant 0, \quad i \in \mathcal{J}^-,$$

où a_i^j est la j-ième composante du vecteur a_i. Le minimum de la fonction économique du problème primal coïncide avec le maximum de celle du problème dual.

6. Problème de programmation quadratique. Le problème de programmation quadratique est de minimiser une fonction quadratique

$$f_0(x) = \frac{1}{2}(x, Cx) + (d, x)$$

sous les contraintes (3.4). Ici C est une matrice $n \times n$ symétrique définie positive et d un n-vecteur.

LEMME 3.7 (sans démonstration.) *Dans le problème de programmation quadratique la borne inférieure ou bien est atteinte ou bien vaut* $-\infty$.

THÉORÈME 3.8. *Pour que le point x_* soit solution du problème de programmation quadratique, il faut et il suffit qu'il existe des nombres u_*^i, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, tels que*

$$Cx_* + d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u_*^i a_i = 0,$$

$$u_*^i = 0 \ \text{si} \ (a_i, x_*) - b_i < 0, \quad i \in \mathcal{J}^-,$$

$$u_*^i \geqslant 0, \quad i \in \mathcal{J}^-.$$

On le démontre par application immédiate du théorème 3.2.

Soit maintenant C strictement définie positive, i.e. il existe un γ tel que $(x, Cx) \geqslant \gamma \| x \|^2$. Elle est alors régulière et admet un inverse, à savoir C^{-1}. Construisons le problème dual

$$\varphi(u) = \inf_{x \in E^n} [f_0(x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i(x)] =$$

$$= \inf_{x \in E^n} \left[\frac{1}{2}(x, Cx) + (d, x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i ((a_i, x) - b_i)\right] =$$

$$= \inf_{x \in E^n} \left[- \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i + \frac{1}{2}(x, Cx) + \left(x, d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i\right)\right].$$

En égalant à zéro les dérivées du second membre trouvons que le minimum est réalisé pour

$$x(u) = -C^{-1}\left(d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i\right).$$

Ceci étant,

$$\varphi(u) = - \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i b_i - \frac{1}{2}\left(d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i, \right.$$
$$\left. C^{-1}\left(d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i\right)\right). \tag{3.15}$$

Le problème dual est donc de maximiser (3.15) avec les contraintes $u^i \geqslant 0$, $i \in \mathcal{J}^-$.

THÉORÈME 3.9. *Si dans le problème de programmation quadratique le minimum est réalisé et la matrice C est strictement définie positive, on a pour ce problème le théorème de Kuhn et Tucker* (3.3) *et le théorème* 3.4. *Ceci étant, les multiplicateurs de Lagrange du problème primal sont solution du problème dual, et si u_* est solution du problème dual, celle du primal s'obtient par la formule*

$$x(u) = -C^{-1}\left(d + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i a_i\right). \tag{3.16}$$

§ 4. Conditions nécessaires de minimum

Le problème général de programmation mathématique consiste à minimiser une fonction $f_0(x)$, $x \in E^n$, sur un ensemble défini par le système d'inégalités et d'égalités

$$f_i(x) \leqslant 0, \quad i \in \mathcal{J}^-, \quad f_i(x) = 0, \quad i \in \mathcal{J}^0, \quad x \in X. \tag{4.1}$$

Ici $\mathcal{J}^-$ et $\mathcal{J}^0$ sont deux ensembles finis d'indices. On supposera partout dans ce paragraphe que $f_i(x)$ sont des fonctions continûment dérivables à gradient $f_i'(x)$. Quant à l'ensemble X, nous nous abstiendrons pour le moment de supposer quoi que ce soit à son sujet.

Ce paragraphe se propose comme tâche principale d'établir des conditions nécessaires imposées au point x_* qui réalise le minimum de $f_0(x)$ sous les contraintes (4.1).

1. Définitions fondamentales.

DÉFINITION 4.1. *Un ensemble D de points vérifiant les contraintes (4.1) s'appelle domaine admissible.*

Cet ensemble sera supposé non vide.

DÉFINITION 4.2. *Une fonction $f_0(x)$ à minimiser sur D s'appelle fonction économique.*

DÉFINITION 4.3. *Un point x_* vérifiant (4.1) tel que*

$$f_0(x_*) \leqslant f_0(x), \quad x \in D,$$

s'appelle point de minimum.

DÉFINITION 4.4. *On dit que x_* est un point de minimum local de $f_0(x)$ sur D s'il possède un voisinage Ω tel que*

$$f(x_*) \leqslant f_0(x), \quad x \in D \cap \Omega.$$

Dans la suite on considère en général le problème de recherche du minimum de $f_0(x)$. On conçoit que maximiser une fonction $f(x)$ sur D se ramène à minimiser $f_0(x) = -f(x)$ sur le même domaine.

2. Conditions nécessaires de minimum.

DÉFINITION 4.5. *Un vecteur $p \in E^n$ définit une direction admissible par rapport à l'ensemble X au point $x_0 \in X$ si, quels que soient les vecteurs $e_i \in E^n$, $i \in \mathcal{J}^0$, et les fonctions $r^i(\lambda)$, $i \in \mathcal{J}^0$, vérifiant la condition*

$$\lim_{\lambda \to 0} \frac{r^i(\lambda)}{\lambda} = 0, \tag{4.2}$$

on a

$$x_0 + \lambda p + \sum_{i \in \mathcal{J}^0} r^i(\lambda) e_i \in X \tag{4.3}$$

pour $\lambda > 0$ suffisamment petits.

Formulons le résultat essentiel qu'on va démontrer dans ce numéro.

THÉORÈME 4.1. *Etant donné x_*, un point de minimum local de $f_0(x)$ sur D, si l'ensemble des directions admissibles par rapport à X en x_* forme un cône convexe $K(x_*)$, il existe des nombres u^0, $u^i \in \mathcal{J}^- \cup \mathcal{J}^0$ tels que*

$$\begin{gathered} u^0 f_0'(x_*) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i'(x_*) \in K^*(x_*), \\ u^i f_i(x_*) = 0, \quad i \in \mathcal{J}^-, \quad u^i \geqslant 0, \quad i = 0, \ i \in \mathcal{J}^-. \end{gathered} \tag{4.4}$$

DÉMONSTRATION. Considérons deux cas.

1) Les vecteurs $f_i'(x_*)$, $i \in \mathcal{J}^0$, sont linéairement dépendants. Il existe alors des nombres u^i, $i \in \mathcal{J}^0$, tels que

$$\sum_{i \in \mathcal{J}^0} u^i f_i'(x_*) = 0.$$

Posons $u^0 = 0$, $u^i = 0$, $i \in \mathcal{J}^-$. Nous constatons la validité de toutes les propositions du théorème.

2) Les vecteurs $f_i'(x_*)$, $i \in \mathcal{J}^0$, sont linéairement indépendants. Il existe alors des vecteurs e_i, $i \in \mathcal{J}^0$, tels que

$$(f_i'(x_*), e_j) = \delta_{ij}, \ i, j \in \mathcal{J}^0,$$

où $\delta_{ij} = 0$ si $i \neq j$ et $\delta_{ii} = 1$.

Soit m le nombre d'indices i de l'ensemble $\mathcal{J}^- \cup \mathcal{J}^0$. Considérons dans E^{m+1} un ensemble Z défini comme suit. Un vecteur z est élément de Z si et seulement si il existe un vecteur $p \in K(x_*)$ tel que

$$z^i = (f_i'(x_*), p) \quad \text{si} \quad f_i(x_*) = 0, \ i \in \mathcal{J}^- \cup \mathcal{J}^0 \text{ ou } i = 0.$$

Les composantes z^i de $z \in Z$ pour lesquelles $f_i(x_*) < 0$ sont arbitraires. Du moment que $K(x_*)$ est un cône convexe, on voit aisément qu'il en est également de Z.

Définissons P. Un vecteur w appartient à P si et seulement si

$$\begin{gathered} w^i < 0 \text{ si } f_i(x_*) = 0, \ i \in \mathcal{J}^- \text{ ou } i = 0, \\ w^i = 0 \text{ si } i \in \mathcal{J}^0. \end{gathered}$$

Les autres composantes de w sont quelconques. P est évidemment un ensemble convexe.

Démontrons que Z et P sont disjoints. Admettons le contraire. Il existe alors un vecteur $p_0 \in K(x_*)$ tel que

$$\begin{gathered} (f_i'(x_*), p_0) < 0, \\ (f_i'(x_*), p_0) < 0 \text{ si } i \in \mathcal{J}^- \text{ et } f_i(x_*) = 0, \\ (f_i'(x_*), p_0) = 0, \ i \in \mathcal{J}^0. \end{gathered} \tag{4.5}$$

Formons un système d'équations en $r^i(\lambda)$, $i \in \mathcal{J}^0$:

$$f_i\left(x_* + \lambda p_0 + \sum_{i \in \mathcal{J}^0} r^i e_i\right) = 0, \quad i \in \mathcal{J}^0. \tag{4.6}$$

Notons

$$g_i(\lambda, r) \equiv f_i\left(x_* + \lambda p_0 + \sum_{i \in \mathcal{J}^0} r^i e_i\right), \quad i \in \mathcal{J}^0.$$

Le système (4.6) est alors équivalent à

$$g_i(\lambda, r) = 0, \quad i \in \mathcal{J}^0, \tag{4.7}$$

qui définit r^i comme des fonctions implicites de λ. $f_i(x)$ étant par hypothèse continûment dérivables, les fonctions $g_i(\lambda, r)$ et r^i sont elles aussi continûment dérivables par rapport à λ. Dans ce cas on calcule facilement que

$$\frac{\partial g_i(0, 0)}{\partial \lambda} = (f_i'(x_*), p_0) = 0, \quad i \in \mathcal{J}^0, \tag{4.8}$$

en vertu de 4.5,

$$\frac{\partial g_i(0, 0)}{\partial r^j} = (f_i'(x_*), e_j) = \delta_{ij}. \tag{4.9}$$

Désignons par $\partial g/\partial r$ la matrice de composantes $\frac{\partial g_i(0, 0)}{\partial r^j}$, i, $j \in \mathcal{J}^0$. D'après le théorème de fonctions implicites, le système (4.7) est résoluble en r pour λ petits si la matrice $\partial g/\partial r$ est régulière. $r(\lambda)$ est dans ce cas une fonction de λ dérivable, $r(0) = 0$ et

$$r'(\lambda) = -\left(\frac{\partial g}{\partial r}\right)^{-1} \frac{\partial g}{\partial \lambda}, \tag{4.10}$$

avec $\frac{\partial g}{\partial \lambda}$ un vecteur de composantes $\frac{\partial g_i(0, 0)}{\partial \lambda}$. Dans le cas examiné

$$\frac{\partial g}{\partial r} = I, \quad \frac{\partial g}{\partial \lambda} = 0, \tag{4.11}$$

où I est la matrice unité. Cela découle de (4.8) et (4.9).

Nous voyons donc que pour λ petits sont définies des fonctions continûment dérivables $r^i(\lambda)$, $i \in \mathcal{J}^0$. De plus

$$\lim_{\lambda \to 0} \frac{r^i(\lambda) - r^i(0)}{\lambda} = \lim_{\lambda \to 0} \frac{r^i(\lambda)}{\lambda} = r^{i\prime}(0) = 0 \tag{4.12}$$

en vertu de (4.10) et (4.11).

Soit maintenant $x(\lambda) = x_* + \lambda p_0 + \sum_{i \in \mathcal{J}^0} r^i(\lambda) e_i$. Par définition de $K(x_*)$ on a alors $x(\lambda) \in X$ pour $\lambda > 0$ petits car $p_0 \in K(x_*)$. Ensuite, $f_i(x(\lambda)) = 0$, $i \in \mathcal{J}^0$, du moment que $r^i(\lambda)$ vérifient (4.6)

par définition. Poursuivons. On a $f_0(x(\lambda)) < f_0(x_*)$ pour $\lambda > 0$ petits. En effet, selon la formule de Taylor

$$f_0(x(\lambda)) = f_0(x_*) + (f_0'(\xi),\ x(\lambda) - x_*),$$

avec ξ un point du segment réunissant x_* à $x(\lambda)$. Aussi

$$\frac{f_0(x(\lambda)) - f_0(x_*)}{\lambda} = (f_0'(\xi),\ p_0) + \sum_{i \in \mathcal{J}^0} \frac{r^i(\lambda)}{\lambda} (f_0'(\xi),\ e_i).$$

Comme, en vertu de (4.5), $(f_0'(x_*), p_0) < 0$ et $\frac{r^i(\lambda)}{\lambda} \to 0$, on a, pour $\lambda > 0$ petits, $\xi \to x_*$ et

$$\frac{f_0(x(\lambda)) - f_0(x_*)}{\lambda} < 0.$$

De même, si $i \in \mathcal{J}^-$ et $f_i(x_*) = 0$, il résulte de (4.5)

$$f_i(x(\lambda)) < 0, \quad i \in \mathcal{J}^-,\ f_i(x_*) = 0.$$

Si $f_i(x_*) < 0$, $i \in \mathcal{J}^-$, alors $f_i(x(\lambda)) < 0$ par continuité.

Ainsi, pour λ positifs faibles, le point $x(\lambda)$ satisfait à toutes les contraintes (4.1) et $f_0(x(\lambda)) < f_0(x_*)$. Or, cela contredit l'hypothèse de x_* point de minimum local.

Vu cette contradiction, les ensembles Z et P ne se coupent donc pas. Etant donné leur convexité, on peut les séparer. Il existe par conséquent des nombres u^0, $u^i \in \mathcal{J}^- \cup \mathcal{J}^0$ non tous nuls tels que

$$u^0 z^0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i z^i \geqslant u^0 w^0 + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i w^i, \quad z \in Z,\ w \in P. \tag{4.13}$$

La structure de Z et P permet de faire certaines conclusions sur u^i. En effet, w^0 peut prendre, par définition de P, une valeur quelconque inférieure à zéro. Il en découle $u^0 \geqslant 0$, sinon le second membre pourrait être aussi grand qu'on le veut, ce qui contredit (4.13). De même

$$u^i \geqslant 0 \text{ si } f_i(x_*) = 0,\ i \in \mathcal{J}^-. \tag{4.14r}$$

Ensuite, si $i \in \mathcal{J}^-$ et $f_i(x_*) < 0$, alors w^i est quelconque. Pour avoir (4.13) il faut donc que

$$u^i = 0 \text{ si } f_i(x_*) < 0,\ i \in \mathcal{J}^-. \tag{4.15}$$

Faisons tendre w de (4.13) vers zéro de façon que $w \in P$; compte tenu de (4.15) et de la définition de Z, il vient

$$u^0 (f_0'(x_*),\ p) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i (f_i'(x_*),\ p) \geqslant 0, \quad p \in K(x_*),$$

ou

$$(u^0 f_0'(x_*) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i'(x_*),\ p) \geqslant 0, \quad p \in K(x_*). \tag{4.16}$$

Les assertions démontrées (4.14), (4.15), (4.16) sont évidemment équivalentes à la proposition du théorème.

CONSÉQUENCE 4.1. *Si $X = E^n$, pour que le point x_* soit minimum local, il faut qu'il existe des nombres u^i non tous nuls tels que*

$$u^0 f_0'(x_*) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i'(x_*) = 0,$$

$$u^0 \geqslant 0, \quad u^i \geqslant 0, \quad i \in \mathcal{J}^-, \quad u^i f_i(x_*) = 0, \quad i \in \mathcal{J}^- \tag{4.17}$$

DÉMONSTRATION. Si $X = E^n$, toute direction p est admissible, i.e. $K(x_*) = E^n$. Le cône $K^*(x_*)$ se compose donc d'un vecteur nul et les relations (4.4) deviennent de suite (4.17).

CONSÉQUENCE 4.2. *Pour que le point x_* réalise le minimum de $f_0(x)$ dans le domaine*

$$x^j \geqslant 0, \quad j \in \mathcal{Y},$$

où $\mathcal{Y}$ est un sous-ensemble des indices 1, 2, . . ., *n, il faut qu'on ait*

$$\frac{\partial f_0(x_*)}{\partial x^j} \geqslant 0 \quad \text{si} \quad x_*^j = 0, \quad j \in \mathcal{Y},$$

$$\frac{\partial f_0(x_*)}{\partial x^j} = 0 \quad \text{si} \quad x_*^j > 0, \quad j \in \mathcal{Y} \quad \text{ou} \quad j \notin \mathcal{Y}. \tag{4.18}$$

DÉMONSTRATION. Les contraintes $x^j \geqslant 0$, $j \in \mathcal{Y}$, peuvent se récrire comme $(-a_j, x) \leqslant 0$, $j \in \mathcal{Y}$, avec a_j un vecteur de composantes $a_j^i = \delta_{ij}$, $i = 1, \ldots, n$. Moyennant la conséquence précédente on établit qu'il existe des nombres u^0 et u^j, $j \in \mathcal{Y}$, non tous nuls tels que

$$u^0 f_0'(x_*) - \sum_{j \in \mathcal{Y}} u^j a_j = 0, \quad u^0, \ u^j \geqslant 0, \quad u^j x_*^j = 0, \quad j \in \mathcal{Y}. \tag{4.19}$$

La première relation (4.19) se récrit

$$u^0 \frac{\partial f_0(x_*)}{\partial x^i} - \sum_{j \in \mathcal{Y}} u^j \delta_{ij} = 0$$

ou

$$u^0 \frac{\partial f_0(x_*)}{\partial x^i} = u^i, \quad i \in \mathcal{Y}, \quad u^0 \frac{\partial f_0(x_*)}{\partial x^i} = 0, \quad i \notin \mathcal{Y} \tag{4.20}$$

Il en résulte $u^0 > 0$ car, dans le cas $u^0 = 0$, il en est également de tous les u^i, ce qui contredit la conséquence 4.1. On admet donc que $u^0 = 1$.

(4.20) et (4.19) entraînent immédiatement la conséquence.

DÉFINITION 4.6. *Dans le problème de minimisation de $f_0(x)$ sous les contraintes* (4.1) *avec $X = E^n$, le point de minimum x_* est*

dit régulier si les gradients $f_i'(x_*)$ *pour des indices i tels que* $i \in \mathcal{J}^- \cup \mathcal{J}^0$, $f_i(x_*) = 0$, *sont linéairement indépendants.*

CONSÉQUENCE 4.3. *Si* x_* *est un point régulier, on pose* $u^0 = 1$ *dans* (4.17) *et les multiplicateurs de Lagrange* u^i, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, *sont définis de façon unique.*

DÉMONSTRATION. En effet, $u^0 > 0$ dans le cas considéré. Car, dans le cas $u^0 = 0$, en vertu de (4.17) les gradients $f_i'(x_*)$ pour lesquels $i \in \mathcal{J}^- \cup \mathcal{J}^0$, $f_i(x_*) = 0$, seraient linéairement dépendants. Ensuite, conformément à (4.17), $u^i = 0$ si $f_i(x_*) < 0$. Aussi la première relation (4.17) avec $u^0 = 1$ fournit le développement

$$f_0'(x_*) = -\sum_{\substack{i \in \mathcal{J}^- \cup \mathcal{J}^0 \\ f_i(x_*)=0}} u^i f_i'(x_*)$$

du vecteur $f_0'(x_*)$ suivant les vecteurs linéairement indépendants $f_i'(x_*)$ et définit donc de façon unique les coefficients u^i.

Supposons maintenant que le problème (4.1) ne possède que les contraintes égalité

$$f_i(x) = 0, \quad i \in \mathcal{J}^0,$$

et $X = E^n$. Si x_* est un point de minimum de $f_0(x)$ pour de telles contraintes et si les gradients $f_i'(x_*)$ sont linéairement indépendants, les conditions nécessaires de minimum (4.17) prennent la forme

$$f_0'(x_*) + \sum_{i \in \mathcal{J}^0} u^i f_i'(x_*) = 0.$$

L'ensemble des vecteurs p vérifiant dans ce cas la condition

$$(f_i'(x_*), p) = 0, \quad i \in \mathcal{J}^0,$$

s'appelle variété tangente à l'ensemble

$$D = \{x : f_i(x) = 0, \ i \in \mathcal{J}^0\}$$

au point x_*.

CONSÉQUENCE 4.4. *Pour que le point* x_* *en lequel* $f_i'(x_*)$, $i \in \mathcal{J}^0$, *sont linéairement indépendantes, réalise le minimum de* $f_0(x)$ *sur l'ensemble* D, *il faut que le gradient* $f_0'(x_*)$ *soit orthogonal à la variété tangente à* D *en* x_*, *i.e. si* p *appartient à la variété tangente, alors* $(f_0'(x_*), p) = 0$. *En d'autres termes, la projection du vecteur* $f_0'(x_*)$ *sur la variété tangente vaut* 0.

DÉMONSTRATION. Si x_* est, sous les hypothèses faites, un point de minimum, alors

$$(f_0'(x_*), p) = -\sum_{i \in \mathcal{J}^0} u^i (f_i'(x_*), p) = 0$$

pour tout p élément de la variété tangente. Inversement, si $(f_0'(x_*), p)$ est nul, quel que soit p appartenant à la variété tangente, il existe la représentation

$$f_0'(x_*) = -\sum_{i \in \mathcal{J}^0} u^i f_i'(x_*),$$

comme cela résulte du lemme 1.6, si l'on décompose chacune des égalités $(f_i'(x_*), p) = 0$ en deux inégalités

$$(f_i'(x_*), p) \geqslant 0, \quad -(f_i'(x_*), p) \geqslant 0.$$

3. Problème de minimax. On demande de minimiser la fonction

$$f(x) = \max_{i=1, \ldots, m} f_i(x), \tag{4.21}$$

où $f_i(x)$ sont des fonctions continûment dérivables et $x \in E^n$. Afin d'appliquer les résultats du numéro précédent ramenons le problème de minimisation de $f(x)$ à un problème équivalent de programmation mathématique, à savoir on constate aisément qu'à condition d'introduire une variable supplémentaire x^{n+1}, x_*, point de minimum de $f(x)$, est également solution du problème : minimiser $g_0(x, x^{n+1}) = x^{n+1}$ dans les contraintes

$$g_i(x, x^{n+1}) \equiv f_i(x) - x^{n+1} \leqslant 0, \quad i = 1, \ldots, m. \tag{4.22}$$

La valeur minimale de $g_0(x, x^{n+1})$ est alors $x_*^{n+1} = f(x_*)$.

Appliquons à (4.22) la conséquence 4.1. Ceci faisant, on tiendra compte du fait qu'on se place maintenant dans l'espace E^{n+1} des variables $x^1, \ldots, x^n, x^{n+1}$, de sorte que les gradients des fonctions $g_i(x, x^{n+1})$ ont la forme

$$g_i'(x, x^{n+1}) = \begin{pmatrix} f_i'(x) \\ -1 \end{pmatrix}, \quad i = 1, \ldots, m, \quad g_0'(x, x^{n+1}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Appliquons la conséquence 4.1 ; on trouve qu'il existe des nombres u^0, u^i, $i = 1, \ldots, m$, non tous nuls tels que

$$\begin{gathered} u^0 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \sum_{i=1}^{m} u^i \begin{pmatrix} f_i'(x_*) \\ -1 \end{pmatrix} = 0, \\ u^i \geqslant 0, \quad i = 0, 1, \ldots, m, \\ u^i (f_i(x_*) - x_*^{n+1}) = u^i (f_i(x_*) - f(x_*)) = 0, \quad i = 1, \ldots, m. \end{gathered} \tag{4.23}$$

La première relation (4.22) donne $u^0 = \sum_{i=1}^{m} u^i$. Vu que $u^i \geqslant 0$ on en tire $u^0 > 0$ parce que, si $u^0 = 0$, il en serait également de tous les u^i. L'homogénéité par rapport à u^i des relations (4.22) permet d'estimer que $u^0 = 1$.

On a donc finalement le

THÉORÈME 4.2. *Pour que x_* soit un point de minimum de $f(x)$ défini par la relation* (4.21), *il faut qu'il existe des nombres u^i, $i = 1, \ldots, m$, tels que*

$$\sum_{i=1}^{m} u^i f_i'(x_*) = 0,$$
$$\sum_{i=1}^{m} u^i = 1, \quad u^i \geqslant 0, \quad i = 1, \ldots, m, \tag{4.24}$$
$$u^i (f_i(x_*) - f(x_*)) = 0, \quad i = 1, \ldots, m.$$

4. Conditions nécessaires du deuxième ordre. Reprenons le problème de minimisation de $f_0(x)$ avec les contraintes (4.1), $X = E^n$. Notons

$$L(x, u) = f_0(x) + \sum_{i \in \mathcal{J}^- \cup \mathcal{J}^0} u^i f_i(x). \tag{4.25}$$

Admettons que le point x_*, solution du problème posé, est régulier (définition 4.6). En vertu de la conséquence 3 du théorème 4.1, la première relation (4.17) s'écrit alors

$$L_x'(x_*, u) = 0. \tag{4.26}$$

Admettons maintenant que toutes les fonctions $f_i(x)$ sont deux fois continûment dérivables, i.e. il existe les matrices continues des dérivées secondes $f_i''(x)$. Cela définit également la matrice des dérivées secondes $L_{xx}''(x_*, u)$ de la fonction $L(x, u)$ par rapport à x.

Le point x_* étant supposé régulier, la relation (4.17) définit les multiplicateurs u^i, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, de façon unique. Notons

$$\mathcal{J}_0^-(x_*) = \{i\colon u^i > 0, \quad i \in \mathcal{J}^-\},$$
$$\mathcal{J}^-(x_*) = \{i\colon f_i(x_*) = 0, \quad i \in \mathcal{J}^-\}.$$

En vertu de (4.17), $\mathcal{J}_0^-(x_*) \subset \mathcal{J}^-(x_*)$. Soit maintenant p vérifiant es inégalités

$$(f_i'(x_*), p) \leqslant 0, \quad i \in \mathcal{J}^-(x_*), \quad i \notin \mathcal{J}_0^-(x_*), \tag{4.27}$$
$$(f_i'(x_*), p) = 0, \quad i \in \mathcal{J}_0^-(x_*) \cup \mathcal{J}^0.$$

Posons

$$\mathcal{J}_p(x_*) = \{i \in \mathcal{J}^-(x_*) \cup \mathcal{J}^0\colon (f_i'(x_*), p) = 0\}. \tag{4.28}$$

La régularité de x_* détermine l'indépendance linéaire des vecteurs $f_i'(x_*)$, $i \in \mathcal{J}_p(x_*)$. On montre donc l'existence d'une fonction $r(\lambda) \in E^n$ telle que

$$f_i(x(\lambda)) = 0, \quad i \in \mathcal{J}_p(x_*), \tag{4.29}$$

où $x(\lambda) = x_* + \lambda p + r(\lambda)$, $\lim\limits_{\lambda \to 0} \frac{r(\lambda)}{\lambda} = 0$. On procède comme dans la démonstration du théorème 4.1.

Si $i \notin \mathcal{J}_p(x_*)$ on a soit $f_i(x_*) < 0$, soit $(f_i'(x_*), p) < 0$, ce qui garantit dans les deux cas l'inégalité $f_i(x(\lambda)) < 0$ pour λ petits. Ainsi, pour λ petits le point $x(\lambda)$ vérifie toutes les contraintes (4.1), $X = E^n$. On en conclut, compte tenu de (4.27)-(4.29),

$$f_0(x(\lambda)) = L(x(\lambda), u).$$

car si $u^i \neq 0$, on a $f_i(x(\lambda)) = 0$ en vertu de (4.29). Par ailleurs, (4.17) entraîne $f_0(x_*) = L(x_*, u)$. Etant donné que $x(\lambda)$ vérifie toutes les contraintes (4.1) et que x_* réalise le minimum de $f_0(x)$ sous ces contraintes, on obtient maintenant pour λ petits

$$L(x(\lambda), u) \geqslant L(x_*, u)$$

Mais en développant $L(x(\lambda), u)$ suivant les puissances de λ aux termes du second ordre infinitésimal près on obtient

$$L(x(\lambda), u) = L(x_*, u) + (L_x'(x_*, u), x(\lambda) - x_*) +$$
$$+ \frac{1}{2}(L_{xx}''(\xi(\lambda), u)(x(\lambda) - x_*), x(\lambda) - x_*) \geqslant L(x_*, u).$$

où $\xi(\lambda)$ est un point du segment joignant x_0 et $x(\lambda)$ de sorte que $\xi(\lambda) \to x_*$ pour $\lambda \to 0$. Utilisons (4.26), nous obtenons

$$\frac{1}{2}\lambda^2 \left(L_{xx}''(\xi(\lambda), u)\left(p + \frac{r(\lambda)}{\lambda}\right), \ p + \frac{r(\lambda)}{\lambda}\right) \geqslant 0.$$

En divisant par λ^2 et en faisant tendre λ vers zéro on a en définitive

$$(L_{xx}''(x_*, u)\, p, p) \geqslant 0.$$

Nous venons de démontrer le

THEOREME 4.3. *Soient $f_i(x)$ des fonctions deux fois continûment dérivables et x_* un point régulier de minimum de $f_0(x)$ dans les contraintes* (4.1), *$X = E^n$. Il existe des nombres u^i, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, tels que*

$$L_x'(x_*, u) = 0, \quad u^i \geqslant 0, \quad i \in \mathcal{J}^-, \quad u^i f_i(x_*) = 0, \quad i \in \mathcal{J}^-,$$

et

$$(L_{xx}''(x_*, u)\, p, p) \geqslant 0$$

pour tous les p vérifiant les inégalités (4.27).

§ 5. Certains renseignements complémentaires

En analyse mathématique on considère la formule de Newton-Leibniz établissant la relation entre une fonction scalaire $f(x)$ et sa dérivée. Cette formule se généralise au cas de fonctions opératorielles.

Si $F(x)$ est une fonction opératorielle dérivable définie sur un

ensemble convexe ouvert $\Omega \in E^n$ *et* x, $x + h \in \Omega$, *alors*

$$F(x+h) - F(x) = \int_0^1 F'(x+\alpha h)\, h d\alpha. \tag{5.1}$$

Pour la démonstration de cette formule (qui est également valable dans le cas d'opérateurs définis dans des espaces fonctionnels) voir, par exemple, A. Kolmogorov et S. Fomine [82].

Voici une autre propriété des fonctions opératorielles.

Si $F(x)$ *est une fonction opératorielle dérivable non linéaire, on a, quels que soient* x, h, $y \in E^n$,

$$(F(x+h) - F(x),\ y) = (F'(x+\theta h)\, h,\ y), \quad 0 \leqslant \theta \leqslant 1. \tag{5.2}$$

C'est la formule de Lagrange pour les opérateurs (ou la formule généralisée de Lagrange). Elle est démontrée (pour des opérateurs de forme plus générale) p. ex. dans la monographie de M. Vainberg [3; 1]. Dans les chapitres qui suivront, nous nous servirons souvent de la formule de Taylor avec le reste sous forme de Lagrange.

Si $f(x)$ est une fonction deux fois continûment dérivable sur un ensemble convexe Ω, on a pour tous les x, $x + h \in \Omega$ et $\alpha \in [0,1]$

$$f(x+\alpha h) - f(x) = \alpha (f'(x+\alpha\theta_1 h),\ h)$$

et

$$f(x+\alpha h) = f(x) + \alpha (f'(x),\ h) + \frac{\alpha^2}{2} (f''(x+\alpha\theta_2 h)\, h,\ h),$$

avec θ_1, $\theta_2 \in [0, 1]$.

COMMENTAIRE

Les propriétés des ensembles et fonctions convexes dans des espaces de dimension finie sont décrites par *S. Karlin* [79], *G. Zoutendijk* [104; 1], *H. Künzi et W. Crelle* [83]. La description la plus exhaustive est contenue dans le livre de *R. Rockafellar* [95]. Les propriétés des ensembles et fonctions convexes dans les espaces fonctionnels sont le sujet d'une étude détaillée due à *N. Dunford et J. Schwartz* [56].

Pour les propriétés spéciales liées à la convexité stricte et forte voir l'article de *E. Lévitine et B. Poliak* [22].

Parmi de nombreux ouvrages consacrés à la théorie des conditions nécessaires d'extrémum citons ceux de *S. Karlin* [79], *G. Zoutendijk* [104; 1], *S. Zoukhovitski et L. Avdééva* [14], *H. Künzi et W. Crelle* [83] qui considèrent les problèmes de programmation linéaire et convexe dans des espaces de dimension finie. Une théorie plus complète des conditions nécessaires du premier ordre dans le cas général est donnée par *A. Doubovitski et A. Miliutine* [11], *L. Neustadt* [90], *H. Halkin et L. Neustadt* [70], *B. Pchénitchny* [29; 1], *M. Hestenes* [72; 2], *K. Arrow, L. Hurwicz et H. Uzawa* [39]. Le livre de Hestenes expose également les conditions nécessaires du deuxième ordre.

Les problèmes de programmation linéaire sont abordés à fond dans *D. Gale* [64], *G. Dantzig* [53; 1], *S. Karlin* [79], *S. Zoukhovitski et L. Avdééva* [14], où on trouve également des algorithmes de calcul et la théorie de la dualité en programmation linéaire. Dans le cas général, la théorie de la dualité en programmation convexe est étudiée par *E. Golstein* [6; 1, 2] et *R. Rockafellar* [95].

Le théorème de fonctions implicites utilisé par les auteurs pour établir les conditions nécessaires d'extrémum est formulé dans *G. Fikhtengoltz* [33].

CHAPITRE II

MÉTHODES DE MINIMISATION DE FONCTIONS SANS CONTRAINTES

Le présent chapitre est consacré au problème de minimisation d'une fonction $f(x)$ définie dans l'espace euclidien n-dimensionnel E^n. Ainsi, x est partout dans ce chapitre un n-vecteur.

On fera appel à des processus itératifs du type

$$x_{k+1} = x_k + \alpha_k p_k, \tag{0.1}$$

où p_k détermine la direction de déplacement à partir du point x_k et α_k est un facteur numérique dont la grandeur donne la longueur du pas dans la direction p_k.

On définit le processus (0.1) en indiquant les procédés de construction du vecteur p_k et de calcul de α_k à chaque itération. La façon dont on construit ce vecteur et définit α_k détermine directement les propriétés du processus: le comportement de la fonction sur les termes de la suite $\{x_k\}$, la convergence de la suite vers la solution, la vitesse de convergence, etc. D'autre part, le volume de calcul et les contraintes imposées à la fonction à minimiser diffèrent selon les procédés de formation de p_k et de α_k.

Voyons les principes qui vont nous guider dans le choix de la direction de déplacement et de la grandeur du pas.

Pour s'approcher du point x_* (dans le cas général, c'est un point en lequel ont lieu — peut-être avec une certaine précision — les conditions nécessaires d'extrémum de $f(x)$), on se déplace naturellement à partir du point x_k dans la direction de la décroissance de la fonction, i.e. dans la direction de descente. Si x_k n'est ni point de minimum ni point stationnaire, il existe une infinité de vecteurs p déterminant la direction de descente à partir de x_k et chacun d'eux est défini par la condition (pour $f(x)$ dérivable)

$$(f'(x_k), p) < 0.$$

Cela découle des considérations suivantes.

Soit $x = x_k + \alpha p$. Moyennant le développement de Taylor de la fonction autour du point x_k (on suppose évidemment que celle-ci

est dérivable un nombre suffisant de fois), on obtient la représentation

$$f(x) = f(x_k) + \alpha (f'_k, p) + \frac{\alpha^2}{2} (f''_{kc} p, p).$$

On a ici $f'_k = f'(x_k)$, $f''_{kc} = f''(x_{kc})$, $x_{kc} = x_k + \theta (x - x_k)$, $\theta \in [0,1]$. Dans ce chapitre nous nous servirons souvent de ces désignations afin d'alléger l'exposé.

Si $(f'_k, p) < 0$, tout au moins pour de petites valeurs du paramètre α, on a $f(x) < f(x_k)$ car le signe du second membre se définit alors par un terme linéaire par rapport à α.

En choisissant de diverses manières la direction de descente et le facteur α_k on trouve divers algorithmes de minimisation.

§ 1. Méthodes du gradient

1. Méthode de la plus grande pente. Le plus simple est de choisir une direction p_k vérifiant la condition $(f'_k, p_k) < 0$ (i.e. une direction de descente de $f(x)$) en posant $p_k = -f'_k$.

Le processus itératif

$$x_{k+1} = x_k - \alpha_k f'(x_k), \ \alpha_k > 0, \ k = 0,1, \ldots, \tag{1.1}$$

ainsi obtenu s'appelle *méthode de la plus grande pente* ou *du gradient*. Il s'écrit en coordonnées

$$x^i_{k+1} = x^i_k - \alpha_k \frac{\partial f(x_k)}{\partial x^i}, \quad i = 1, 2, \ldots, n.$$

La méthode utilisant la plus grande pente est actuellement un procédé de minimisation parmi les plus connus. Sa popularité s'explique essentiellement par sa simplicité relative et par la possibilité d'aborder une très vaste classe de fonctions.

Etudions les propriétés de l'algorithme (1.1) et commençons par la technique de choix du facteur scalaire α_k.

1) Choisissons une valeur de α (la même pour toutes les itérations) et fixons le point $x = x_k - \alpha f'_k$.

2) Calculons $f(x) = f(x_k - \alpha f'_k)$.

3) Testons l'inégalité

$$f(x) - f(x_k) \leqslant \varepsilon \alpha (f'_k, p_k), \tag{1.2}$$

où $0 < \varepsilon < 1$ est une constante quelconque (la même pour tous les $k = 0, 1, \ldots$).

4) Si (1.2) a lieu, la valeur de α est prise pour valeur cherchée : $\alpha_k = \alpha$. Dans le cas contraire, on fractionne α (en le multipliant par un nombre arbitraire $\lambda < 1$) tant que l'inégalité en question n'est pas satisfaite.

Ce choix de α_k demande une motivation : il faut établir les conditions d'existence des valeurs non nulles du paramètre α pour lesquelles on a (1.2). On démontre notamment le

THÉORÈME 1.1. *Si une fonction $f(x)$ est minorée, son gradient $f'(x)$ vérifie la condition de Lipschitz*

$$\| f'(x) - f'(y)^1 \| \leqslant R \| x - y \| \tag{1.3}$$

quels que soient $x, y \in E^n$, et le choix de α_k s'effectue de façon décrite, alors, quel que soit le point initial x_0, on a pour le processus (1.1) $\| f'_k \| \to 0$ *lorsque* $k \to \infty$

DÉMONSTRATION. Selon le théorème de la moyenne

$$f(x) - f(x_k) = (f'(x_{kc}),\ x - x_k). \tag{1.4}$$

où $x_{kc} = x_k + \theta (x - x_k)$, $\theta \in [0,1]$. L'indice « kc »(« c ») désignera partout dans la suite un point intermédiaire au segment correspondant.

L'égalité (1.4) se transforme en

$$f(x) - f_k = (f'_k,\ x - x_k) + (f'_{kc} - f'_k,\ x - x_k).$$

D'où, vu que $x - x_k = -\alpha f'_k$ et utilisant (1.3) :

$$f(x) - f_k \leqslant -\alpha (f'_k,\ f'_k) + \alpha R \| x_{kc} - x_k \| \| f'_k \| \leqslant$$
$$\leqslant -\alpha \| f'_k \|^2 + \alpha R \| x - x_k \| \| f'_k \| = \alpha \| f'_k \|^2 (-1 + \alpha R)$$

Cette estimation montre l'existence de valeurs $\alpha \neq 0$ telles que l'inégalité (1.2) soit valable : il suffit de choisir α de sorte que $-1 + \alpha R \leqslant -\varepsilon$, ce qui est parfaitement possible parce que R est borné et $0 < \varepsilon < 1$. Par conséquent, on a (1.2) tout au moins pour $\alpha \leqslant \frac{1-\varepsilon}{R}$. En choisissant α_k par l'algorithme cité on obtient donc

$$f_{k+1} - f_k \leqslant -\varepsilon \alpha_k \| f'_k \|^2, \tag{1.5}$$

i.e. $f_{k+1} - f_k < 0$ pour tout k (si $\| f'_k \| \neq 0$). La fonction étant minorée par hypothèse, la dernière inégalité entraîne pour $k \to \infty$

$$f_{k+1} - f_k \to 0. \tag{1.6}$$

Il résulte de (1.5)

$$\| f'_k \|^2 \leqslant \frac{f_k - f_{k+1}}{\varepsilon \alpha_k} \tag{1.7}$$

Notons à présent que l'algorithme décrit de choix de α_k garantit $\alpha_k \geqslant \bar{\alpha} > 0$ quel que soit k, le rôle de $\bar{\alpha}$ étant tenu par toute constante au plus égale à $\frac{1-\varepsilon}{R}$ puisque l'inégalité (1.2) (ou (1.5)) est nécessairement remplie (nous l'avons déjà dit) pour $\alpha = \frac{1-\varepsilon}{R}$.

Compte tenu de cette remarque, il découle des conditions (1.6) et (1.7) que $\| f'_k \| \to 0$ pour $k \to \infty$, ce qui démontre le théorème.

La classe de fonctions satisfaisant aux exigences du théorème 1.1 est fort vaste. De telles fonctions peuvent en général être dépourvues de points de minimum, elles peuvent présenter des minima locaux, des points-selles, etc. Le théorème 1.1 montre que la méthode du gradient garantit la convergence pour la fonction soit vers la borne inférieure $\inf_x f(x)$, soit vers la valeur de la fonction en un point stationnaire. On a également convergence de la suite $\{x_k\}$ vers un point stationnaire (quand il existe), mais déterminer sa vitesse sous les conditions imposées à la fonction par le théorème 1.1 est chose difficile. Dans des conditions suffisamment sévères imposées à la régularité et à la convexité de la fonction, non seulement on démontre la convergence de la suite $\{x_k\}$, mais aussi on apprécie la vitesse de convergence.

THÉORÈME 1.2. *Soit $f(x)$ une fonction deux fois continûment dérivable, sa matrice des dérivées secondes vérifiant les conditions*

$$m \| y \|^2 \leqslant (f''(x)\, y, y) \leqslant M \| y \|^2, \quad M \geqslant m > 0, \tag{1.8}$$

quels que soient $x, y \in E^n$, et la suite $\{x_k\}$ étant construite par la méthode (1.1), où α_k est choisi de façon décrite plus haut. Quel que soit le point initial x_0, on a alors $x_k \to x_*$, $f(x_k) \to f(x_*)$, où x_* est un point de minimum (unique) de $f(x)$.

La vitesse de convergence est évaluée par

$$f_k - f_* \leqslant q^k [f_0 - f_*], \quad \| x_k - x_* \| \leqslant C q^{k/2}, \tag{1.9}$$
$$C < \infty, \ 0 < q < 1.$$

DÉMONSTRATION. L'existence et l'unicité du minimum de $f(x)$ dans les conditions du théorème découlent des résultats du lemme I.3.2. Il nous reste donc à démontrer la convergence de $\{x_k\}$ vers x_* et à obtenir les estimations (1.9). Etablissons d'abord la deuxième de ces dernières. Moyennant la formule de Taylor, on a

$$f(x_*) = f(x) + (f'(x), x_* - x) + \frac{1}{2} (f''(x_c)(x_* - x), x_* - x),$$

d'où, compte tenu de (1.8)

$$f(x) - f(x_*) \leqslant (f'(x), x - x_*) - \frac{m}{2} \| x - x_* \|^2 \leqslant$$
$$\leqslant \| f'(x) \| \, \| x - x_* \| - \frac{m}{2} \| x - x_* \|^2. \tag{1.10}$$

Par ailleurs, du moment que $f'(x_*) = 0$,

$$f(x) - f(x_*) = \frac{1}{2} (f''(x_{c_1})\ (x - x_*), x - x_*),$$

c'est pourquoi, en vertu de (1.8),

$$\frac{m}{2}\|x-x_*\|^2 \leqslant f(x)-f(x_*) \leqslant \frac{M}{2}\|x-x_*\|^2. \tag{1.11}$$

Utilisons la première inégalité (1.11) et (1.10); nous établissons que

$$\|x-x_*\| \leqslant \frac{\|f'(x)\|}{m}, \tag{1.12}$$

et la dernière inégalité (1.11) entraîne

$$\|x-x_*\|^2 \geqslant \frac{2}{M}[f(x)-f(x_*)].$$

Compte tenu des estimations obtenues, (1.10) s'écrit comme

$$f(x)-f(x_*) \leqslant \frac{\|f'(x)\|^2}{m} - \frac{m}{M}[f(x)-f(x_*)],$$

d'où

$$\|f'(x)\|^2 \geqslant m\left(1+\frac{m}{M}\right)[f(x)-f(x_*)]. \tag{1.13}$$

Introduisant cette estimation dans l'inégalité (1.5) nous trouvons

$$f_{k+1}-f_k \leqslant -\varepsilon\alpha_k m\left(1+\frac{m}{M}\right)(f_k-f_*). \tag{1.14}$$

Sous les hypothèses du théorème

$$f(x)-f(x_k)=(f'_k,\ x-x_k)+\frac{1}{2}(f''_{kc}(x-x_k),\ x-x_k)=$$
$$=-\alpha\|f'_k\|^2+\frac{\alpha^2}{2}(f''_{kc}f'_k,\ f'_k) \leqslant -\alpha\left(1-\frac{\alpha M}{2}\right)\|f'_k\|^2.$$

Il s'ensuit que l'inégalité (1.2) a nécessairement lieu si $1-\frac{\alpha M}{2} \geqslant \varepsilon$, i.e. $\alpha \leqslant \bar{\alpha} = \frac{2(1-\varepsilon)}{M}$. Ceci étant, il résulte de (1.14)

$$f_{k+1}-f_* \leqslant \left[1-\varepsilon\alpha_k m\left(1+\frac{m}{M}\right)\right](f_k-f_*) \leqslant q(f_k-f_*),$$

où $q=1-\varepsilon\bar{\alpha}m\left(1+\frac{m}{M}\right)<1$, i.e.

$$(f_k-f_*) \leqslant q^k(f_0-f_*). \tag{1.15}$$

Puisque $\bar{\alpha}=\frac{2(1-\varepsilon)}{M}$, on a

$$q=1-\frac{2\varepsilon(1-\varepsilon)m}{M}\left(1+\frac{m}{M}\right),$$

ce qui implique que la valeur minimale $q_{\min}$ de la raison de la progression est atteinte pour $\varepsilon = \frac{1}{2}$, de plus

$$q_{\min} = 1 - \frac{m}{2M}\left(1 + \frac{m}{M}\right).$$

Il serait donc bon de poser $\varepsilon = 1/2$ dans la condition (1.2). L'estimation (1.15) conjointement avec la première estimation (1.11) permet d'établir la convergence de la suite $\{x_k\}$ vers le point de minimum et d'évaluer sa vitesse :

$$\| x_k - x_* \| \leqslant \left(\frac{2}{m}\right)^{1/2} (f_k - f_*)^{1/2} \leqslant \left(\frac{2}{m}\right)^{1/2} (f_0 - f_*)^{1/2} q^{k/2} \leqslant C q^{k/2}.$$

Le théorème est démontré.

Quand nous analysons la démonstration ci-dessus, nous voyons que l'estimation (1.15) a été en fait obtenue à l'aide des seules conditions (1.2) et (1.13). On conclut donc que la classe de fonctions vérifiant cette estimation est en réalité beaucoup plus vaste que celle des fonctions satisfaisant aux conditions (1.8), à savoir on a (1.15) pour toutes les fonctions remplissant les conditions du théorème 1.1 et la condition

$$\| f'(x) \|^2 \geqslant \delta\, [f(x) - f_*], \quad \delta > 0.$$

Dans ce cas la démonstration de (1.15) est en fait indépendante de l'existence du minimum : on estime que $f_* = \inf f(x)$ sans se demander si la borne inférieure est atteinte ou non. Il convient tout de même de souligner que les fonctions de cette classe présentent un point de minimum (pas forcément unique), que la suite $\{x_k\}$ converge vers un point x_* et que la vitesse de convergence est évaluée par la deuxième inégalité (1.9).

En effet, utilisant (1.1) et (1.7) on obtient

$$\| x_{k+1} - x_k \|^2 = \alpha_k^2 \| f'_k \|^2 \leqslant \frac{\alpha_k}{\varepsilon} (f_k - f_{k+1}) \leqslant$$

$$\leqslant \frac{\alpha_k}{\varepsilon} (f_k - f_*) \leqslant \frac{\alpha_{\max}}{\varepsilon} q^k [f_0 - f_*] \leqslant C_1 q^k.$$

Ici $\alpha_{\max}$ est la valeur maximale du paramètre à partir de laquelle on commence le fractionnement. Ceci étant, on a pour tout $m > k$

$$\| x_m - x_k \| \leqslant \sum_{i=k}^{m-1} \| x_{i+1} - x_i \| \leqslant C_1^{1/2} \sum_{i=k}^{m-1} q^{i/2} \leqslant C_1^{1/2} \frac{q^{k/2}}{1 - q^{1/2}},$$

d'où $\| x_m - x_k \| \to 0$ pour $k \to \infty$, i.e. sa suite $\{x_k\}$ converge (vers un point de minimum x_*) ; de plus

$$\| x_* - x_k \| = \lim_{m \to \infty} \| x_m - x_k \| \leqslant C_1^{1/2} \frac{q^{k/2}}{1 - q^{1/2}} = C_2 q^{k/2}.$$

2. Certaines variantes de la méthode de la plus grande pente. Le procédé décrit de choix du paramètre α_k dans le processus (1.1), procédé lié à la vérification de l'inégalité (1.2), n'est pas le seul possible. Nous allons examiner plusieurs autres techniques de choix de α_k dont chacune est associée à une méthode du gradient. En démontrant les théorèmes 1.1 et 1.2 on a établi que (1.2) a forcément lieu pour $\alpha \leqslant \frac{1-\varepsilon}{R}$ (théorème 1.1) ou pour $\alpha \leqslant \frac{2(1-\varepsilon)}{M}$ (théorème 1.2). C'est justement cette circonstance qui a permis de démontrer la proposition sur les propriétés de la méthode (1.1) en choisissant α_k par la condition de validité de (1.2). Connaissant les constantes R ou M caractérisant la fonction à minimiser $f(x)$, dans la méthode (1.1) on peut prendre à l'avance $\alpha_k \equiv \bar{\alpha}$, où $0 < \bar{\alpha} \leqslant \frac{1-\varepsilon}{R}$ ou $0 < \bar{\alpha} \leqslant \frac{2(1-\varepsilon)}{M}$, et les théorèmes 1.1 et 1.2 restent en vigueur. Dans cette version de la méthode du gradient on arrive à préciser la grandeur de la raison q dans les estimations de la vitesse de convergence (1.9).

THÉORÈME 1.3. *Si $f(x)$ vérifie les conditions du théorème 1.2 et si dans la méthode (1.1) $\alpha_k = \alpha$, $0 < \alpha < \frac{2}{M}$, la vitesse de convergence de la suite $\{x_k\}$ s'évalue par*

$$\| x_k - x_* \| \leqslant q^k \| x_0 - x_* \|,$$
$$q = \max\{|1-\alpha m|, |1-\alpha M|\},$$

la valeur minimale $q_{\min} = \frac{M-m}{M+m}$ étant atteinte pour $\alpha = \frac{2}{M+m}$.

DÉMONSTRATION. On a

$$\| x_{k+1} - x_* \|^2 = (x_k - \alpha f'_k - x_*, x_{k+1} - x_*) =$$
$$= (x_k - x_* - \alpha (f'_k - f'_*), x_{k+1} - x_*).$$

Utilisant la formule de Lagrange pour les opérateurs (I.5.2) nous obtenons

$$(\alpha (f'_k - f'_*), x_{k+1} - x_*) = (\alpha f''_{kc} (x_k - x_*), x_{k+1} - x_*).$$

Vu cette égalité

$$\| x_{k+1} - x_* \|^2 = ((I - \alpha f''_{kc})(x_k - x_*), x_{k+1} - x_*) \leqslant$$
$$\leqslant \| I - \alpha f''_{kc} \| \, \| x_k - x_* \| \, \| x_{k+1} - x_* \|,$$

i.e. 128

$$\| x_{k+1} - x_* \| \leqslant \| I - \alpha f''_{kc} \| \, \| x_k - x_* \| = q \| x_k - x_* \|.$$

En vertu des conditions (1.8),

$$q = \| I - \alpha f''_{kc} \| = \max\{|1-\alpha m|, |1-\alpha M|\}$$

Sur l'intervalle $(0, 2/M)$ la fonction linéaire $1 - \alpha M$ change manifestement de signe, et la valeur $q_{\min}(\alpha)$ est donc atteinte quand $1 - \alpha m = -(1 - \alpha M)$, i.e. $\alpha = \frac{2}{M+m}$, auquel cas on a évidemment $q_{\min} = \frac{M-m}{M+m}$, c.q.f d.

Notons que lorsque $\alpha = \frac{2}{M+m}$, la première estimation (1.9) est améliorée comme suit :

$$f_{k+1} - f_* \leqslant \left(\frac{M-m}{M+m}\right)^2 (f_k - f_*). \tag{1.16}$$

Indiquons un autre procédé de choix de la longueur du pas. On peut choisir α_k à partir de la condition de minimum de la fonction dans la direction du déplacement, i.e. ce α_k doit garantir la condition

$$f(x_k - \alpha_k f'_k) = \min_{\alpha \geqslant 0} f(x_k - \alpha f'_k). \tag{1.17}$$

Avec ce choix du pas, on conserve tous les résultats relatifs aux propriétés de la méthode (1.1) (voir plus haut) ; de plus la vitesse de convergence s'évalue de façon plus précise.

Démontrons une affirmation analogue au théorème 1.1.

THÉORÈME 1.4. *Si $f(x)$ vérifie les hypothèses du théorème* 1.1 *et si dans la méthode* (1.1) *α_k est choisi à partir de la condition* (1.17), *alors $\| f'_k \| \to 0$ pour $k \to \infty$, quel que soit le point initial x_0.*

DÉMONSTRATION Comme dans le théorème 1.1, on obtient l'estimation

$$f(x) - f(x_k) = -\alpha \| f'_k \|^2 - \alpha (f'_{kc} - f'_k, f'_k) \leqslant$$
$$\leqslant -\alpha \| f'_k \|^2 + \alpha^2 R \| f'_k \|^2.$$

La fonction $\varphi(\alpha) = -\alpha \| f'_k \|^2 + \alpha^2 R \| f'_k \|^2$ atteint son minimum pour $\alpha_{\min} = \frac{1}{2R}$, $\varphi(\alpha_{\min})$ étant égal à $-\frac{\| f'_k \|^2}{4R}$. Puisque $\alpha^2 R \| f'_k \|^2$ majore $-\alpha (f'_{kc} - f'_k, f'_k)$, il est clair que la valeur de α_k satisfaisant à la condition (1.17) est au moins $\alpha_{\min}$ et

$$f_{k+1} - f_k \leqslant -\frac{\| f'_k \|^2}{4R} \tag{1.18}$$

D'où, à force de raisonner comme dans le théorème 1.1, $\| f'_k \| \to 0$, c.q.f.d.

Dans cette méthode du gradient à pas choisi à partir de la condition (1.17) les estimations (1.9) se démontrent comme dans le théorème 1.2, à cette différence près qu'on utilise l'expression (1.13) dans l'estimation $f_{k+1} - f_k \leqslant -\frac{1}{2M} \| f'_k \|^2$ qui s'obtient de même que (1.18). Nous mènerons cependant la démonstration en nous appuyant

sur les résultats du théorème 1.3. Nous obtiendrons ainsi une valeur plus exacte de q, raison de la progression.

Soit $\tilde{x}_{k+1} = x_k - \frac{2}{M+m} f'_k$; on a alors l'estimation (1.16):

$$f(\tilde{x}_{k+1}) - f(x_*) \leqslant \left(\frac{M-m}{M+m}\right)^2 (f_k - f_*).$$

Si l'on choisit x_{k+1} moyennant la condition de minimum dans la direction de déplacement, alors

$$f(x_{k+1}) - f(x_*) \leqslant f(\tilde{x}_{k+1}) - f(x_*) \leqslant \left(\frac{M-m}{M+m}\right)^2 (f_k - f_*) \leqslant$$
$$\leqslant \left(\frac{M-m}{M+m}\right)^{2(k+1)} (f_0 - f_*).$$

Faisons recours aux estimations (1.11), il vient

$$\| x_{k+1} - x_* \|^2 \leqslant \frac{2}{m} (f_{k+1} - f_*) \leqslant \frac{2}{m} \left(\frac{M-m}{M+m}\right)^{2(k+1)} \frac{M}{2} \| x_0 - x_* \|^2$$

ou, finalement,

$$\| x_{k+1} - x_* \| \leqslant C \left(\frac{M-m}{M+m}\right)^{k+1},$$

où $C = \left(\frac{M}{m}\right)^{1/2} \| x_0 - x_* \|$.

On a donc le

THÉORÈME 1.5. *Si $f(x)$ vérifie les hypothèses du théorème 1.2 et si dans la méthode (1.1) α_k est choisi à partir de la condition (1.17), la suite $\{x_k\}$ converge en progression géométrique de raison $q = \frac{M-m}{M+m}$ vers le point de minimum.*

C'est justement à la variante de (1.1) avec le pas choisi par la condition de minimum dans la direction de déplacement qu'on donne souvent le nom de méthode de la plus grande pente.

3. D'autres méthodes du gradient. Soit $F(x)$ une matrice symétrique arbitraire vérifiant les conditions

$$\rho \| y \|^2 \leqslant (F(x) y, y) \leqslant R \| y \|^2, \ \rho > 0, \tag{1.19}$$

pour tous les $x, y \in E^n$. Si l'on choisit un vecteur $p = -F(x) f'(x)$, alors $(f'(x), p) = -(f', Ff') \leqslant -\rho \| f' \|^2 < 0$ sous la condition $\| f'(x) \| \neq 0$. Ainsi, le vecteur $p = -F(x) f'(x)$ définit une direction de descente de $f(x)$. Fort de ce résultat, on construit pour minimiser $f(x)$ le processus itératif

$$x_{k+1} = x_k - \alpha_k F_k f'(x_k), \ \alpha_k > 0, \ k = 0, 1, \ldots,$$

avec $\{F_k\}$ une suite de matrices arbitraires vérifiant (1.19). Afin de se conformer à l'exposé ultérieur de ce chapitre (plus précisé-

ment, pour ne pas contredire le symbolisme adopté), on considérera le processus

$$x_{k+1} = x_k - \alpha_k F_k^{-1} f'_k, \ \alpha_k > 0, \tag{1.20}$$

où figure l'inverse de la matrice F_k. Cela est absolument sans importance car si F_k vérifie les conditions (1.19), son inverse satisfait (voir lemme I.2.9) à

$$m_1 \| y \|^2 \leqslant (F^{-1}y, y) \leqslant M_1 \| y \|^2, \ m_1 = \frac{\rho}{R^2} > 0, \quad M_1 = \frac{1}{\rho} \tag{1.21}$$

et donc

$$(f'_k, p_k) = -(f'_k, F_k^{-1} f'_k) \leqslant -m_1 \| f'_k \|^2 \leqslant 0. \tag{1.22}$$

A diverses suites $\{F_k^{-1}\}$ correspondent des processus itératifs différents.

Quant à l'idée de la méthode (1.20), son étude n'offre aucun élément nouveau par rapport à la méthode « pure » du gradient (1.1). Tous les résultats obtenus pour celle-ci se conservent pour la méthode (1.20) sous les mêmes conditions imposées à la fonction à minimiser et le même procédé de choix du pas. La seule différence, qui est d'ailleurs insignifiante, tient à la démonstration des affirmations correspondantes. On conçoit également que les valeurs numériques des paramètres de (1.20) diffèrent de celles des paramètres analogues du processus (1.1). C'est le cas en particulier de la grandeur de q (raison de la progression) dans les estimations de la vitesse de convergence.

Arrêtons-nous sur des résultats de la méthode (1.20) dont nous nous servirons dans la suite.

THÉORÈME 1.6. *Les résultats du théorème* 1.2 *restent valables pour la méthode* (1.20).

DÉMONSTRATION. Si $x = x_k + \alpha p_k$, où $p_k = -F_k^{-1} f'_k$, alors

$$f(x) - f(x_k) = \alpha (f'_k, p_k) + \frac{\alpha^2}{2} (f''_{kc} p_k, p_k) \leqslant$$
$$\leqslant \alpha (f'_k, p_k) \left(1 + \frac{\alpha M \| p_k \|^2}{2 (f'_k, p_k)}\right).$$

Mais compte tenu de (1.19)

$$(f'_k, p_k) = -(F_k p_k, p_k) \leqslant -\rho \| p_k \|^2. \tag{1.23}$$

Par conséquent,

$$f(x) - f(x_k) \leqslant \alpha (f'_k, p_k) \left(1 - \frac{\alpha}{2} \frac{M}{\rho}\right).$$

Il s'ensuit nécessairement l'inégalité (1.2) si $1-\frac{\alpha M}{2\rho}\geqslant\varepsilon$, i.e. $\alpha\leqslant\bar{\alpha}=$ $=\frac{2(1-\varepsilon)}{M}\rho$. Nous avons donc justifié le choix de α_k.

Puisque $(f'_k, p_k)<0$ pour $\| f'_k \|\neq 0$, la condition

$$f_{k+1}-f_k\leqslant\varepsilon\alpha_k\,(f'_k, p_k) \tag{1.24}$$

entraîne $f_{k+1}<f_k$. Utilisons (1.24) et le fait que $f(x)$ est minoré et établissons, comme $\| f'_k \|\to 0$ dans le théorème 1.1, que (f'_k, p_k) tend vers zéro pour $k\to\infty$. Cela signifie, en vertu de (1.22), que $\| f'_k \|\to 0$. D'où, vu la convexité forte de $f(x)$, la convergence de (1.20) vers la solution x_*. Pour évaluer la vitesse de convergence $f_k\to f_*$, $x_k\to x_*$, représentons l'inégalité (1.24) moyennant (1.22) comme $f_{k+1}-f_k\leqslant-\varepsilon\alpha_k m_1\| f'_k \|^2$. Exprimant ce $\| f'_k \|$ à l'aide de l'inégalité (1.13) et répétant mot à mot les raisonnements du théorème 1.2, nous établissons la validité des estimations (1.9) pour la méthode (1.20). La raison de la progression est alors

$$q=1-\varepsilon\bar{\alpha}m_1m\left(1+\frac{m}{M}\right)=1-\varepsilon\,\frac{2(1-\varepsilon)\rho}{M}\,m_1m\left(1+\frac{m}{M}\right),$$

et elle atteint sa valeur minimale pour $\varepsilon=1/2$:

$$q_{\min}=1-\frac{\rho^2m}{2R^2M}\left(1+\frac{m}{M}\right).$$

Le théorème est démontré.

La démonstration ci-dessus entraîne qu'il y a toujours convergence du processus (1.20) si on pose $\alpha_k=\alpha$, $0<\alpha<\frac{2}{M}\rho$ (variante à pas constant). Les mêmes raisonnements que dans le théorème 1.3 nous conduisent à l'estimation

$$\| x_{k+1}-x_* \|\leqslant\| I-\alpha F_k^{-1}f''_{kc} \|\,\| x_k-x_* \|.$$

Il est cependant impossible d'en déduire, à l'instar du théorème 1.3, l'estimation de q parce que la matrice $F_k^{-1}f''_{kc}$ n'est en général pas définie positive (cette dernière propriété n'a lieu que pour F_k^{-1} et $f''(x_{kc})$ permutables).

On considère une méthode (1.20) avec un pas choisi à partir de la condition de minimum de $f(x)$ dans la direction de déplacement.

THÉORÈME 1.7. *Si $f(x)$ vérifie les conditions du théorème 1.2 et si, dans la méthode (1.20), le paramètre α_k est choisi par la condition*

$$f(x_k+\alpha_kp_k)=\min_{\alpha\geqslant 0} f(x_k+\alpha p_k),$$

alors la suite $\{x_k\}$ converge en progression géométrique vers le point de minimum.

Adoptons le schéma de démonstration suivant. En utilisant le développement taylorien au second ordre de la fonction autour du point x_k et en raisonnant comme pour le théorème 1.4, on obtient l'estimation

$$f_{k+1} - f_k \leqslant -\frac{1}{2}\frac{(f'_k, p_k)^2}{M \| p_k \|^2}.$$

En vertu de (1.22) et (1.23) elle est équivalente à

$$f_{k+1} - f_k \leqslant -\frac{1}{2}\frac{m_1 \| f'_k \|^2}{M}.$$

Représentons $\| f'_k \|$ moyennant l'inégalité (1.13) et répétons textuellement les développements du théorème 1.2. On ne réussit pas ici à améliorer la valeur de q parce que cette valeur est forcément supérieure à celle de la méthode de la plus grande pente.

4. Analyse qualitative des méthodes. Comparons les méthodes du gradient que nous venons d'examiner et exposons certaines considérations sur la qualité de ces algorithmes, i.e. sur leur efficacité dans la recherche du minimum.

Nous avons étudié trois versions de la méthode (1.1) qui diffèrent par le choix du pas et dont les propriétés sont voisines. Ces méthodes sont utilisées pour minimiser les fonctions d'une même classe et la vitesse de convergence est (quand on réussit à l'évaluer) quasi identique elle aussi. En résolvant le problème, il faut donc recourir à la version la moins laborieuse. Le coût de chaque itération des variantes en question ne diffère évidemment qu'à cause d'une définition différente du paramètre α_k. La variante à pas constant $\alpha_k = \alpha$ exige le moins de calculs par itération (on ne calcule alors que le gradient $f'(x_k)$). Or, ce choix de α_k est pratiquement impossible dans la plupart des problèmes du fait qu'on est d'habitude dans l'ignorance des valeurs des constantes R, M caractérisant la fonction.

Comparons le coût des procédés de choix du pas liés avec la vérification des conditions (1.2) et (1.17). Nous avons établi que dans le cas où $f(x)$ remplit certaines conditions (théorèmes 1.1, 1.2), l'inégalité (1.2) a manifestement lieu pour des α_k suffisamment petits au moins (définis par la grandeur des constantes R, M). Aussi, quelle que soit la valeur initiale $\alpha_{\max}$ à partir de laquelle on commence à vérifier l'inégalité (1.2), celle-ci a lieu au bout d'un nombre fini de fractionnements du paramètre, i.e. le choix d'un α_k nécessaire exige qu'on calcule la fonction un nombre fini de fois. Quant au choix de α_k à partir de la condition (1.17), c'est en général une procédure en dimension infinie. En fait, on définit le point de minimum dans la direction de déplacement en calculant la fonction un nombre fini de fois. Il est clair qu'une solution plus ou moins exacte d'un problème de minimisation à une variable demande un volume de calcul plus important que la vérification de l'inégalité (1.2). On

voit donc qu'on préférera le procédé de choix du pas lié avec cette vérification.

Tout ce que nous avons dit reste vrai pour la méthode (1.20).

Il est inutile de rappeler au lecteur que nos raisonnements ne se basent que sur les propriétés les plus générales de la fonction à minimiser et des algorithmes étudiés et que nous ignorons ce que telle fonction concrète a de spécifique. Nos recommandations n'ont donc pas force de loi. Nous prions le lecteur de se le rappeler tout au long de cet ouvrage.

Passons à la question de l'efficacité des méthodes du gradient. Dans le cas de fonctions régulières et convexes (qui sont suffisamment bonnes du point de vue de la résolution du problème de minimisation), les méthodes du gradient convergent en progression géométrique vers le minimum. La grandeur de la raison de la progression dépend, en particulier pour des fonctions fortement convexes, des valeurs propres extrêmes M et m de la matrice des dérivées secondes de $f(x)$. La raison q ne sera suffisamment petite que si m et M diffèrent peu, i.e. si la matrice $f''(x)$ est bien conditionnée. Les méthodes du gradient convergent dans ce cas très rapidement. La pratique du calcul numérique connaît cependant très peu de tels problèmes. On a en général à minimiser des fonctions ayant la matrice $f''(x)$ mal conditionnée $\left(\frac{m}{M} \ll 1\right)$. Plus le rapport m/M est faible, plus la raison de la progression q est proche de l'unité et plus les méthodes du gradient convergent lentement. On peut interpréter géométriquement ce fait. Avec la diminution du rapport m/M les surfaces de niveau de la fonction à minimiser (i.e. les surfaces $f(x) = C$) s'allongent et en la plupart des points la direction du gradient $f'(x)$ s'écarte de plus en plus de la direction de déplacement vers le point de minimum. C'est là la cause du ralentissement de la convergence. L'exemple le plus tangible est en l'occurrence une fonction quadratique strictement convexe $f(x)$ dans l'espace E^2, p.ex. $f = \frac{1}{2}\left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right)$. La matrice des dérivées secondes de cette fonction a tous ses éléments constants, ses surfaces de niveau sont des ellipses d'équation $\frac{1}{2}\left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right) = C$ et le point de minimum coïncide avec le centre des ellipses. Les valeurs propres de la matrice sont $1/a^2$ et $1/b^2$. Plus le rapport a^2/b^2 diffère de l'unité, plus les lignes de niveau sont allongées le long de l'un des axes OX ou OY et plus de pas il faut faire dans la direction de l'opposé du gradient à partir d'un point arbitraire (x_0, y_0) pour arriver dans un voisinage suffisamment restreint du point de minimum.

Les méthodes du gradient étant à convergence lente, il est impossible d'aborder à leur aide les problèmes complexes de minimisation puisque même avec des machines modernes rapides cette résolu-

tion demande trop de temps. Aussi continue-t-on d'élaborer des techniques à convergence plus rapide et combine-t-on souvent les méthodes du gradient avec d'autres plus puissantes, et cela au départ même, quand le point x_k est loin du minimum et on arrive, en suivant l'opposé du gradient, à faire décroître sensiblement la fonction. Et tout de même nous tenons à souligner une fois de plus les avantages incontestables des procédés utilisant le gradient — leur simplicité et la possibilité d'utiliser des fonctions de nature fort diverse.

§ 2. Méthode de Newton à pas variable

1. Construction de la méthode. Les méthodes du gradient déterminent la direction de déplacement moyennant le seul terme linéaire du développement taylorien de la fonction, i.e. l'approximation la plus grossière de celle-ci.

Soit à minimiser une fonction $f(x)$ strictement convexe et suffisamment régulière.

Considérons la fonction

$$\psi(x) = f(y) + (f'(y), x-y) + \frac{1}{2}(f''(y)(x-y), x-y),$$

approximation quadratique de $f(x)$ au voisinage d'un point y. Vu la stricte convexité de $f(x)$, $\psi(x)$ est, on s'en convainc sans peine, strictement convexe elle aussi et atteint donc son minimum en un seul point, le vecteur $p = \tilde{y} - y$ minimisant $\psi(x)$ étant défini par la formule $p = -(f''(y))^{-1} f'(y)$. La direction p est une direction de descente de $f(x)$ parce que $(f'(y), p) = -(f''(y)p, p) < 0$ en vertu de la convexité de $f(x)$. Dans un voisinage restreint de y la fonction quadratique $\psi(x)$ approxime la fonction à minimiser beaucoup plus exactement que ne le fait une fonction linéaire. On s'attend donc naturellement (du moins si le point y se trouve dans un voisinage suffisamment restreint de la solution x_*) à ce que le déplacement à partir de y dans la direction $p = -(f''(y))^{-1} f'(y)$ permette de réaliser une décroissance plus importante de la fonction et d'obtenir une meilleure approximation de la solution que le déplacement dans la direction $-f'(y)$ utilisée dans la méthode du gradient. Ces raisonnements autorisent à parler d'une plus grande efficacité du processus itératif générant les approximations successives de la solution du problème de minimisation de $f(x)$:

$$x_{k+1} = x_k - \alpha_k (f''_k)^{-1} f'_k, \quad \alpha_k > 0, \quad k = 0, 1, \ldots, \tag{2.1}$$

devant la méthode de la plus grande pente, i.e. avec l'algorithme (2.1) la convergence $x_k \to x_*$, $f(x_k) \to f(x_*)$ sera plus apide que celle de la méthode du gradient.

La méthode (2.1) *sera appelée méthode de Newton à pas variable ou méthode de Newton généralisée.*

La méthode de Newton classique correspond au cas $\alpha_k = 1$.

A condition de noter $\varphi_{ij}(x_k)$, $i, j = 1, 2, \ldots, n$, i étant l'indice de la ligne, les éléments de la matrice $(f''_k)^{-1}$, la méthode (2.1) s'écrit en coordonnées

$$x_{k+1}^i = x_k^i - \alpha_k \sum_{j=1}^{n} \varphi_{ij} \frac{\partial f(x_k)}{\partial x^j}, \quad i = 1, \ldots, n.$$

Cette méthode peut également se mettre sous forme

$$f''_k p_k = -f'_k, \quad x_{k+1} = x_k + \alpha_k p_k$$

ou, en coordonnées,

$$\sum_{j=1}^{n} \frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} p_k^j = -\frac{\partial f(x_k)}{\partial x^i},$$

$$x_{k+1}^i = x_k^i + \alpha_k p_k^i, \quad i = 1, \ldots, n.$$

Pour définir le vecteur p_k on peut donc, au lieu d'inverser la matrice $f''(x_k)$, résoudre un système d'équations linéaires.

Nous allons étudier deux variantes de la méthode de Newton à pas variable, où le paramètre α est choisi de façon différente. Décrivons le premier procédé de choix.

1) Posons $\alpha = 1$ et trouvons le point $x = x_k + \alpha p_k$.
2) Calculons $f(x) = f(x_k + \alpha p_k)$.
3) Vérifions l'inégalité

$$f(x) - f(x_k) \leqslant \varepsilon\alpha(f'_k, p_k), \quad 0 < \varepsilon < \frac{1}{2}. \tag{2.2}$$

4) Si (2.2) a lieu, on prend $\alpha = 1$ pour valeur cherchée : $\alpha_k = 1$. Dans le cas contraire, on partage α jusqu'à ce qu'on ait l'inégalité en question.

Nous appellerons choix de α_k à partir de la condition (2.2) *le choix décrit de la valeur du paramètre α_k*. Il s'agit en l'occurrence d'un procédé analogue à celui de la méthode de la plus grande pente lié à la vérification de l'inégalité (1.2).

Dans la deuxième version de (2.1) la valeur de α_k doit fournir le minimum de la fonction dans la direction de déplacement

$$f(x_k - \alpha_k (f''_k)^{-1} f'_k) = \min_{\alpha \geqslant 0} f(x_k - \alpha (f''_k)^{-1} f'_k). \tag{2.3}$$

2. Théorèmes sur les propriétés de la méthode. Il résulte de la formule (2.1) que la méthode de Newton n'est valable que pour des fonctions dont la matrice des dérivées secondes possède une inverse, la matrice $(f''_k)^{-1}$ devant être bornée (cette dernière propriété sera mise en évidence par la suite). Ces exigences sont remplies par les

fonctions deux fois continûment dérivables fortement convexes. Tout le long du paragraphe nous supposerons donc que $f(x)$ vérifie les conditions

$$m \| y \|^2 \leqslant (f''(x)\, y,\ y) \leqslant M \| y \|^2, \quad m > 0, \tag{2.4}$$

quels que soient $x, y \in E^n$. Rappelons que ces fonctions possèdent un minorant et un seul point de minimum x_*.

THÉORÈME 2.1. *Si l'on minimise une fonction $f(x)$ vérifiant les conditions* (2.4) *par la méthode* (2.1) *où le paramètre α_k est choisi à partir de la condition* (2.2), *alors, quel que soit le choix du point initial x_0, la convergence de la suite $\{x_k\}$ vers le point de minimum est superlinéaire*:

$$\| x_{N+l} - x_* \| \leqslant C \lambda_N \ldots \lambda_{N+l}. \tag{2.5}$$

Ici N, $C < \infty$, $\lambda_{N+l} < 1$ pour tout $l \geqslant 0$, $\lambda_i \to 0$ quand $i \to \infty$.

DÉMONSTRATION. La méthode (2.1) peut s'interpréter comme le processus du type gradient (1.20) si on estime que $F_k^{-1} = (f_k'')^{-1}$. La matrice f_k'' jouissant des propriétés nécessaires, la convergence de (2.1) vers la solution découle des résultats généraux de convergence des méthodes du gradient (théorème 1.6).

Etablissons l'estimation (2.5) et commençons par remarquer que

$$(f_k',\ p_k) = -(f_k'' p_k,\ p_k) \leqslant -m \| p_k \|^2. \tag{2.6}$$

Du moment que $(f_k',\ p_k) < 0$ et $(f_k',\ p_k) \to 0$ (théorème 1.6), (2.6) entraîne que $\| p_k \| \to 0$ pour $k \to \infty$. Montrons qu'à partir d'une certaine itération on a $\alpha_k \equiv 1$ dans la méthode (2.1). Utilisons le développement de Taylor et l'expression (2.6), il vient

$$f_{k+1} - f_k = \alpha_k (f_k',\ p_k) + \frac{\alpha_k^2}{2} (f_k'' p_k,\ p_k) + \frac{\alpha_k^2}{2} ((f_{kc}'' - f_k'')\, p_k,\ p_k) \leqslant$$

$$\leqslant \alpha_k (f_k',\ p_k) \left(1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{\| f_{kc}'' - f_k'' \| \, \| p_k \|^2}{m \| p_k \|^2} \right).$$

Ici $x_{kc} = x_k + \theta (x_{k+1} - x_k)$, $\theta \in [0, 1]$. Puisque $\| x_k - x_* \| \to 0$, on a, pour k tendant vers l'infini,

$$\| f_{kc}'' - f_k'' \| \leqslant \| f_{kc}'' - f_*'' \| + \| f_*'' - f_k'' \| \to 0,$$

en vertu de la continuité de la fonction (opératorielle) $f''(x)$. Il existe donc pour toute constante $0 < \varepsilon < \frac{1}{2}$ un nombre $N_0(\varepsilon)$ tel que, lorsque $k \geqslant N_0(\varepsilon)$, la condition

$$1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{\| f_{kc}'' - f_k'' \|}{m} \geqslant \varepsilon$$

est vérifiée pour $\alpha_k = 1$. Cela signifie qu'il en est également de l'inégalité (2.2). Ainsi, sous les hypothèses du théorème, le choix

utilisé du pas garantit qu'à partir d'une certaine itération la méthode (2.1) se réalise avec le pas unité, i.e. elle devient la méthode de Newton usuelle. On peut maintenant évaluer la vitesse de convergence de (2.1):

$$(x_{k+1} - x_*,\ x_{k+1} - x_*) = (x_k - x_* - (f''_k)^{-1} f'_k,\ x_{k+1} - x_*).$$

En vertu de la formule de Lagrange pour les opérateurs

$$((f''_k)^{-1} f'_k,\ x_{k+1} - x_*) = ((f''_k)^{-1} (f'_k - f'_*),\ x_{k+1} - x_*) = \\ = ((f''_k)^{-1} f''_{kc} (x_k - x_*),\ x_{k+1} - x_*).$$

Ici $x_{kc} = x_k + \theta (x_k - x_*)$, $\theta \in [0, 1]$. Par conséquent,

$$\| x_{k+1} - x_* \|^2 = ((I - (f''_k)^{-1} f''_{kc}) (x_k - x_*),\ x_{k+1} - x_*) = \\ = ((f''_k)^{-1} (f''_k - f''_{kc}) (x_k - x_*),\ x_{k+1} - x_*) \leqslant \\ \leqslant \frac{1}{m} \| f''_k - f''_{kc} \| \, \| x_k - x_* \| \, \| x_{k+1} - x_* \|,$$

ou

$$\| x_{k+1} - x_* \| \leqslant \lambda_k \| x_k - x_* \|, \tag{2.7}$$

où $\lambda_k = \frac{1}{m} \| f''_k - f''_{kc} \|$. Puisque $\| f''_k - f''_{kc} \| \to 0$, on peut indiquer un nombre N tel que, pour $k = N + l$, $l = 0, 1, \ldots$, on ait $\lambda_{N+l} < 1$ et $\lambda_{N+l} \to 0$ quand $l \to \infty$. En posant $\| x_N - x_* \| = C$ et en tenant compte des remarques faites, on a l'estimation (2.5). Le théorème est démontré.

Supposons que la matrice $f''(x)$ satisfait, en plus de (2.4), à la condition de Lipschitz

$$\| f''(x) - f''(y) \| \leqslant R \| x - y \|, \quad x, y \in E^n. \tag{2.8}$$

Dans ce cas, dans l'estimation (2.7)

$$\lambda_k = \frac{1}{m} \| f''_k - f''_{kc} \| \leqslant \frac{R}{m} \| x_k - x_* \|$$

et donc

$$\| x_{k+1} - x_* \| \leqslant \frac{R}{m} \| x_k - x_* \|^2. \tag{2.9}$$

On a donc le

THÉORÈME 2.2. *Si la fonction $f(x)$ est telle qu'on ait* (2.4) *et* (2.8), *quel que soit le point initial x_0, la convergence de la suite* (2.1), *où les valeurs de α_k sont choisies à partir de la condition* (2.2), *vers la solution est quadratique, i.e. on a l'estimation* (2.9).

L'estimation (2.9) peut se mettre sous une autre forme. Notons $\mu_k = \frac{R}{m} \| x_k - x_* \|$. Il existe un nombre L tel que pour $k = L + l$, $l = 0, 1, \ldots$, on ait $\mu_{L+l} < 1$. De plus

$$\frac{R}{m} \| x_{L+l} - x_* \| \leqslant \left(\frac{R}{m} \| x_{L+l-1} - x_* \| \right)^2 \leqslant \ldots \leqslant \left(\frac{R}{m} \| x_L - x_* \| \right)^{2^l}.$$

On écrit en définitive

$$\| x_{L+l} - x_* \| \leqslant \frac{m}{R} \mu_L^{2^l}.$$

Voyons la variante de (2.1) avec choix du pas à partir de la condition (2.3). La convergence de la suite $\{x_k\}$ vers la solution découle dans ce cas des résultats généraux sur la convergence des méthodes du gradient (théorème 1.7). Comme dans la version avec α_k choisi par (2.2), la convergence est superlinéaire si on a (2.4) et quadratique si on a de plus la condition (2.8). On le démontre de la façon suivante.

Soit $\tilde{x}_{k+1} = x_k - (f_k'')^{-1} f_k'$ et $x_{k+1} = x_k - \alpha_k (f_k'')^{-1} f_k'$, avec α_k choisi moyennant la condition (2.3). Utilisons les estimations (1.11), il vient

$$\frac{m}{2} \| x_{k+1} - x_* \|^2 \leqslant f_{k+1} - f_* \leqslant f(\tilde{x}_{k+1}) - f_* \leqslant \frac{M}{2} \| \tilde{x}_{k+1} - x_* \|^2.$$

En vertu de (2.7), $\| \tilde{x}_{k+1} - x_* \| \leqslant \lambda_k \| x_k - x_* \|$, $\lambda_k \to 0$ pour $k \to \infty$. Si les conditions (2.4) sont remplies, on a donc

$$\| x_{k+1} - x_* \| \leqslant \left(\frac{M}{m} \right)^{1/2} \lambda_k \| x_k - x_* \| = \gamma_k \| x_k - x_* \|, \qquad (2.10)$$

où $\gamma_k = \left(\frac{M}{m} \right)^{1/2} \lambda_k \to 0$ pour $k \to \infty$. Si (2.8) est valable, alors $\lambda_k \leqslant \frac{R}{m} \| x_k - x_* \|$ et

$$\| x_{k+1} - x_* \| \leqslant \left(\frac{M}{m} \right)^{1/2} \frac{R}{m} \| x_k - x_* \|^2. \qquad (2.11)$$

3. Versions de la méthode de Newton généralisée. Prenons d'abord l'algorithme où les approximations successives sont construites par la formule

$$x_{k+1} = x_k - \alpha_k (f_0'')^{-1} f_k', \quad \alpha_k \geqslant 0. \qquad (2.12)$$

Dans cette méthode $p_k = -(f_0'')^{-1} f_k'$, i.e. on construit les directions de descente à l'aide de la même matrice $(f_0'')^{-1}$. La méthode (2.12) est un cas particulier de l'algorithme (1.20) ($F_k^{-1} = (f_0'')^{-1}$), et on affirme donc que, quel que soit le point initial x_0, la suite (2.12) converge vers la solution en progression géométrique, le pas pouvant être choisi à partir de (2.2) ou de (2.3) (théorèmes 1.6 et 1.7). Mais le choix de l'approximation initiale x_0 influe essentiellement

sur la grandeur de la raison q de la progression, i.e. la vitesse de convergence réelle. En effet, vu que $(f'_k, p_k) = -(f'_k, f''_0 f'_k)$ pour la méthode (2.12), on obtient moyennant la formule de Taylor l'estimation (comme dans le théorème 2.1):

$$f_{k+1} - f_k \leqslant \alpha_k (f'_k, p_k) \left(1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{\| f''_{kc} - f''_0 \|}{m}\right). \qquad (2.13)$$

Ici $x_{kc} = x_k + \theta (x_{k+1} - x_k)$, $\vartheta \in [0, 1]$. Si $x_0 \to x_*$, on a, en vertu de la continuité de la matrice des dérivées secondes,

$$\max_{x \in S} \| f''(x) - f''(x_0) \| \to 0$$

($S = \{x : f(x) \leqslant f(x_0)\}$). Plus le point initial x_0 est proche de x_* et plus la valeur α_k vérifiant (2.2) doit donc être grande, i.e. plus le pas avec lequel on effectue le processus (2.12) s'avère important si la longueur du pas est choisie à partir de la condition (2.2). En particulier, il existe pour toute constante $0 < \varepsilon < \frac{1}{2}$ une constante $\rho(\varepsilon)$ telle que si l'approximation initiale x_0 est prise dans une sphère S de rayon ρ, alors

$$\frac{1}{2} - \frac{1}{2} \frac{\max\limits_{x \in S} \| f''(x) - f''_0 \|}{m} \geqslant \varepsilon.$$

Cela signifie par suite de (2.13) qu'en choisissant une approximation initiale suffisamment voisine du point x_*, l'inégalité (2.2) est juste pour $\alpha_k = 1$, i.e. le processus (2.12) converge avec le pas unité. De même que pour le théorème 2.1 on obtient l'estimation

$$\| x_{k+1} - x_* \| \leqslant \| (f''_0)^{-1} \| \, \| f''_0 - f''_{kc} \| \, \| x_k - x_* \| \leqslant q \, \| x_k - x_* \|. \qquad (2.14)$$

Ici $q = \frac{1}{m} \max\limits_{x \in S} \| f''_0 - f''(x) \|$. Il en résulte justement que la grandeur de q dépend du choix du point initial x_0, q diminuant avec l'écart entre x_0 et x_*.

Dans la variante de (2.12) avec le pas choisi à partir de la condition de minimum dans la direction de déplacement, on établit, compte tenu de (2.14) et à force de raisonner comme pour les estimations (2.10) et (2.11), que

$$\| x_{k+1} - x_* \| \leqslant \left(\frac{M}{m}\right)^{1/2} q \, \| x_k - x_* \| = \tilde{q} \, \| x_k - x_* \|,$$

où $\tilde{q} = \left(\frac{M}{m}\right)^{1/2} \frac{1}{m} \max\limits_{x \in S} \| f''_0 - f''_{kc} \| \to 0$ si $x_0 \to x_*$.

Et voici une autre méthode de Newton modifiée.

Soit $k = \xi t + i$, $\xi = 0, 1, \ldots, i = 0, 1, \ldots, t-1$, $t \geqslant 1$ étant un entier quelconque. On construit alors le processus itératif

$$x_{\xi t+i+1} = x_{\xi t+i} - \alpha_{\xi t+i} (f''_{\xi t})^{-1} f'_{\xi t+i}, \quad \alpha_k \geqslant 0,$$

ou, avec les notations initiales,

$$x_{k+1} = x_k - \alpha_k (f''_{\xi t})^{-1} f'_k, \quad \alpha_k \geqslant 0. \tag{2.15}$$

Il s'agit d'un algorithme intermédiaire entre (2.1), où on construit à chaque fois le vecteur p_k moyennant une nouvelle matrice $(f''_k)^{-1}$, et (2.12) qui détermine la direction de déplacement par une même $(f''_0)^{-1}$. Dans la méthode (2.15), la matrice est renouvelée tous les t pas. Comme (2.1) et (2.12), l'algorithme (2.15) s'interprète comme variante de la méthode du gradient (1.20), et pour divers choix du pas sa convergence résulte des théorèmes 1.6 et 1.7.

Voyons quelle est la convergence de la méthode (2.15) avec le pas choisi à partir de la condition (2.2) en supposant que la fonction $f(x)$ vérifie (2.4) et (2.8).

Appliquant la formule de Taylor on obtient

$$f_{k+1} - f_k \leqslant \alpha_k (f'_k, p_k) \left(1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2} \frac{\| f''_{kc} - f''_{\xi t} \|}{m}\right).$$

La convergence du processus fait que $\| x_{kc} - x_{\xi t} \| = \| x_{(\xi t+i)c} - x_{\xi t} \| \leqslant \| x_{\xi t} - x_{\xi t+1} \| + \ldots + \| x_{\xi t+i} - x_{\xi t+i+1} \| \to 0$ pour $k \to \infty$. Par suite, $\| f''_{kc} - f''_{\xi t} \| \to 0$. Vu cette circonstance on montre par les mêmes raisonnements que pour le théorème 2.1 qu'à partir d'une certaine itération la méthode (2.15) se réalise avec le pas unité : $\alpha_k = 1$. En vertu du théorème 2.2, on a alors l'estimation

$$\| x_{\xi t+1} - x_* \| \leqslant \frac{R}{m} \| x_{\xi t} - x_* \|^2 \tag{2.16}$$

quels que soient $\xi \geqslant L$, L un entier positif. Comme dans le théorème 2.1 il vient ensuite l'estimation

$$\| x_{\xi t+2} - x_* \| = \| x_{\xi t+1} - (f''_{\xi t})^{-1} f'_{\xi t+1} - x_* \| \leqslant$$
$$\leqslant \| (f''_{\xi t})^{-1} \| \, \| f''_{\xi t} - f''_{(\xi t+1)c} \| \, \| x_{\xi t+1} - x_* \|.$$

Ici $x_{(\xi t+1)c} = x_{\xi t+1} + \theta (x_* - x_{\xi t+1})$, $\theta \in [0, 1]$. Conformément à (2.8), $\| f''_{\xi t} - f''_{(\xi t+1)c} \| \leqslant \| f''_{\xi t} - f''_* \| + \| f''_* - f''_{(\xi t+1)c} \| \leqslant R (\| x_{\xi t} - x_* \| + \| x_{\xi t+1} - x_* \|)$. Compte tenu de (2.16) on trouve

$$\| x_{\xi t+2} - x_* \| \leqslant \frac{R}{m} (\| x_{\xi t} - x_* \| + \| x_{\xi t+1} - x_* \|) \| x_{\xi t+1} - x_* \| \leqslant$$
$$\leqslant \frac{R^2}{m^2} \| x_{\xi t} - x_* \|^3 \left(1 + \frac{R}{m} \| x_{\xi t} - x_* \|\right),$$

i.e.

$$\| x_{\xi t+2} - x_* \| \leqslant C_2 \| x_{\xi t} - x_* \|^3, \quad C_2 < \infty.$$

Supposons que pour un $2 < j < t - 1$ on a l'estimation

$$\| x_{\xi t+j} - x_* \| \leqslant C_j \| x_{\xi t} - x_* \|^{j+1}, \quad C_j < \infty.$$

Alors

$$\begin{aligned}\| x_{\xi t+j+1} - x_* \| &= \| x_{\xi t+j} - (f''_{\xi t})^{-1} f'_{\xi t+j} - x_* \| \leqslant \\ &\leqslant \| (f''_{\xi t})^{-1} \| \, \| f''_{\xi t} - f''_{(\xi t+j)c} \| \, \| x_{\xi t+j} - x_* \| \leqslant \\ &\leqslant \frac{R}{m} (\| x_{\xi t} - x_* \| + \| x_{\xi t+j} - x_* \|) \, \| x_{\xi t+j} - x_* \| \leqslant \\ &\leqslant C_{j+1} \| x_{\xi t} - x_* \|^{j+2}, \quad C_{j+1} = \frac{R}{m} C_j (1 + C_j \| x_{\xi t} - x_* \|^j).\end{aligned}$$

La vitesse de convergence de la méthode (2.15) s'évalue donc par

$$\| x_{(\xi+1)t} - x_* \| \leqslant C \| x_{\xi t} - x_* \|^{t+1}. \tag{2.17}$$

Cette estimation signifie que la convergence de la suite $\{x_{\xi t}\}$ vers la solution est d'ordre $t + 1$.

4. Discussion des propriétés de la méthode de Newton. Nous avons établi que la méthode de Newton à pas variable converge vers la solution indépendamment du choix du point initial x_0 et que cette convergence est superlinéaire ou quadratique selon les contraintes satisfaites par la fonction $f(x)$.

Le fait de converger à partir de n'importe quelle approximation initiale constitue un avantage sérieux de la méthode (2.1) sur la méthode de Newton classique où la convergence n'est garantie qu'en présence d'une approximation initiale suffisamment bonne (i.e. suffisamment voisine de la solution du problème). Par ailleurs, la vérification des conditions sous lesquelles une approximation initiale donnée garantit la convergence s'avère fort difficile dans la méthode classique de Newton parce qu'elle exige qu'on sache de la fonction des choses qu'on ignore en général (par exemple, la valeur des constantes m, M).

La comparaison de deux procédés d'ajustement du pas où on teste respectivement (2.2) et (2.3) montre que la première technique coûte moins sur le plan du calcul de la fonction (en particulier, à partir d'une certaine itération elle exige que celle-ci ne soit calculée qu'une seule fois, vu que $\alpha_k = 1$) et garantit une convergence au moins aussi rapide que la deuxième.

S'agissant de la minimisation de fonctions convexes, la méthode de Newton assure une convergence des approximations successives vers la solution infiniment plus rapide que les méthodes du gradient. Si l'on entend donc par efficacité d'une méthode sa vitesse de convergence, la méthode de Newton est, comme nous l'avons supposé au début du paragraphe, beaucoup plus efficace que celles du gradient. Or, il serait plus exact d'apprécier l'efficacité d'un algorithme selon le volume de calcul qu'il exige pour résoudre tel problème

avec une précision désirée. L'efficacité d'une méthode se mesure donc au nombre d'itérations nécessaires pour arriver à la solution et au nombre d'opérations arithmétiques par itération.

Dans la méthode de Newton ce dernier nombre est en général beaucoup plus élevé que dans les techniques utilisant le gradient. La cause en est la nécessité de calculer et d'inverser la matrice des dérivées secondes. Par contre, un problème est résolu par cette méthode (avec une précision assez bonne) en un nombre d'itérations des dizaines et des centaines de fois moindre que par les méthodes du gradient, d'où son efficacité sensiblement plus grande. Pour certains problèmes son coût par itération peut néanmoins s'avérer catastrophique, et la cause en est toujours le calcul de la matrice des dérivées secondes $f''(x)$ (en général, la difficulté majeure de problèmes d'extrémum est justement le calcul de $f''(x)$ et non pas son inversion). Plus bas nous aborderons de tels problèmes. Dans ce cas on peut utiliser les versions examinées de la méthode de Newton dont l'une se contente d'un seul calcul et d'une seule inversion de $f''(x)$ et l'autre ne l'exige qu'au bout d'un nombre fini d'itérations. Ceci étant, pour une approximation initiale, assez bonne, on a une convergence rapide. Mais on ne tranche pas la question du coût (qui peut en général croître) par le recours à ces versions. On veut donc construire des méthodes de minimisation qui convergent à peu près comme la méthode de Newton tout en demandant à chaque itération un volume de calcul sensiblement moindre. On dispose de plusieurs méthodes pareilles construites par le jeu de considérations différentes. Elles s'avèrent en général plus puissantes que la méthode de Newton et leur vogue ne cesse donc de grandir. L'étude de tels algorithmes fait l'objet des paragraphes suivants de ce chapitre.

§ 3. Méthodes des directions duales

1. Sur le choix du schéma des méthodes. Nous avons noté dans le paragraphe précédent la difficulté majeure de la méthode de Newton qu'est le calcul de la matrice des dérivées secondes de la fonction à minimiser. Pour être plus efficace, un algorithme doit donc se passer de cette opération tout en conservant la vitesse de convergence de la méthode de Newton.

Une question se pose: ne vaudrait-il pas mieux générer les approximations successives de la solution en construisant des directions p_k voisines de celles de la méthode de Newton à l'aide de la seule dérivée première de la fonction à minimiser?

Les dérivées premières et secondes de $f(x)$ sont liées par la formule de Taylor pour les opérateurs (c'est justement le cas du gradient $f'(x)$):

$$f'(y) - f'(x) = f''(x)(y - x) + \omega(x, y - x), \tag{3.1}$$

où $\| \omega(x, y - x) \| = o(\| y - x \|)$.

L'égalité (3.1) fait penser que si on calcule les dérivées $f'(x)$ en des points $x_1, \ldots, x_{n+1}$ arbitraires mais voisins et si on définit une matrice carrée A $n \times n$ moyennant le système d'équations (vectorielles)

$$f'(x_{i+1}) - f'(x_i) = A(x_{i+1} - x_i), \quad i = 1, \ldots, n \tag{3.2}$$

(en supposant certes les vecteurs $x_{i+1} - x_i$, $i = 1, \ldots, n$, linéairement indépendants), A sera proche de la matrice des dérivées secondes calculée en un point x_i quelconque.

En effet, on a pour tout i, en vertu de (3.1),

$$f'(x_{i+1}) - f'(x_i) = f''(x_i)(x_{i+1} - x_i) + \omega(x_i, x_{i+1} - x_i)$$

et donc, compte tenu de (3.2),

$$A(x_{i+1} - x_i) = f''(x_i)(x_{i+1} - x_i) + \omega(x_i, x_{i+1} - x_i), \quad i = 1, \ldots, n.$$

Ce système d'équations se récrit comme

$$A(x_{i+1} - x_i) = f''_j(x_{i+1} - x_i) + (f''_i - f''_j)(x_{i+1} - x_i) + \\ + \omega(x_i, x_{i+1} - x_i), \quad i = 1, \ldots, n, \quad 1 \leqslant j \leqslant n. \tag{3.3}$$

Si la matrice $f''(x)$ est régulière et continue, étant donné la proximité des points x_i, la somme de deux derniers termes du second membre de chaque équation du système (3.3) doit être sensiblement inférieure au premier terme, i.e.

$$A(x_{i+1} - x_i) \approx f''(x_j)(x_{i+1} - x_i), \quad i = 1, \ldots, n,$$

ce qui atteste en général la proximité des matrices A et f''_j, $j = 1, \ldots, n$. On conçoit sans peine comment construire, les considérations ci-dessus aidant, les processus itératifs de minimisation. Si $\{x_k\}$ est une suite formée de façon quelconque qui converge vers le point de minimum de $f(x)$, dans un voisinage suffisamment petit de ce minimum les points $x_k, x_{k-1}, \ldots, x_{k-n}$ sont proches les uns des autres. En définissant la matrice A_k par le système d'équations

$$A_k(x_{k-i} - x_{k-i-1}) = f'(x_{k-i}) - f'(x_{k-i-1}), \quad i = 0, 1, \ldots, n-1,$$

on peut construire la $(k+1)$-ième approximation par la formule

$$x_{k+1} = x_k - \alpha_k A_k^{-1} f'_k, \quad \alpha_k > 0. \tag{3.4}$$

Si A_k s'avère suffisamment proche de f''_k, $p_k = -A_k^{-1} f'_k$ est alors voisine de $-(f''_k)^{-1} f'_k$ (direction de déplacement de la méthode de Newton) et c'est donc une direction de descente. En définissant de façon analogue les matrices $A_{k+1}, A_{k+2}, \ldots$, leur proximité des matrices $f''_{k+1}, f''_{k+2}, \ldots$ fait que le processus (3.4) doit posséder des propriétés proches de celles de la méthode de Newton. D'autre part,

on n'a pas besoin dans ce processus de calculer les dérivées secondes de la fonction.

Ces considérations nous autorisent donc à dire qu'il est possible de construire toute une classe de processus de descente dont la convergence est superlinéaire et qui se passent du calcul des dérivées secondes de la fonction. Nous avons appelé ces processus méthodes des directions duales pour des raisons qui seront expliquées lorsque nous examinerons les procédés de calcul de la matrice A_k^{-1} et du vecteur p_k.

Occupons-nous maintenant à justifier rigoureusement les méthodes du type (3.4).

2. Justification des méthodes. Supposons que $f(x)$ admet des dérivées premières et secondes continues. Soit $\{x_k\}$ une suite infinie à qui nous faisons correspondre une suite $\{y_k\}$ définie par la formule

$$y_k = x_k + r_k, \tag{3.5}$$

avec r_k des vecteurs vérifiant les conditions suivantes:

1) si Δ_k est un déterminant de colonnes constituées par les vecteurs $\frac{r_k}{\| r_k \|}, \ldots, \frac{r_{k-n+1}}{\| r_{k-n+1} \|}$, alors, quel que soit $k \geqslant n-1$, on a $|\Delta_k| \geqslant \varepsilon$, ε étant un nombre positif petit;

2) $\| r_k \| \to 0$ pour $k \to \infty$. Ce sont les seules contraintes sur le choix des vecteurs r_k.

La première exigence ci-dessus est en fait la condition d'indépendance linéaire des r_k.

LEMME 3.1. *Etant donné une suite bornée* $\{x_k\}$, $\| x_{k+1} - x_k \| \to 0$ *pour* $k \to \infty$, *et une matrice* A_k *définie, pour tout* $k \geqslant n-1$, *par le système d'équations*

$$A_k r_{k-i} = e_{k-i}, \quad i = 0, 1, \ldots, n-1, \tag{3.6}$$

où $e_{k-i} = f'(y_{k-i}) - f'(x_{k-i})$, r_k, y_k *sont les éléments de la suite* (3.5), *on a*

$$\lim_{k \to \infty} \| A_k - f''(x_k) \| = 0.$$

DÉMONSTRATION. Utilisons la formule (I.5.1) et représentons l'opérateur $f'(x)$ par sa dérivée:

$$f'(y_{k-i}) - f'(x_{k-i}) = \int_0^1 f''(x_{k-i} + \tau(y_{k-i} - x_{k-i}))\, r_{k-i}\, d\tau =$$

$$= \int_0^1 f''(x_{k-i})\, r_{k-i}\, d\tau + \int_0^1 [f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i})]\, r_{k-i}\, d\tau =$$

$$= f''(x_{k-i})\, r_{k-i} + \int_0^1 [f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i})]\, r_{k-i}\, d\tau.$$

Compte tenu de ces égalités,

$$(A_k - f''(x_k))\, r_{k-i} = (f''(x_{k-i}) - f''(x_k))\, r_{k-i} + \\ + \int_0^1 [f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i})]\, r_{k-i}\, d\tau.$$

Notons $A_k - f''_k = B_k$, il vient

$$\| B_k r_{k-i} \| \leqslant \| f''_{k-i} - f''_k \| \, \| r_{k-i} \| + \\ + \sup_{0 \leqslant \tau \leqslant 1} \| f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i}) \| \, \| r_{k-i} \|. \tag{3.7}$$

La suite $\{x_k\}$ étant bornée, on a pour tout k $x_k \in Q$, ($Q \subset E^n$ est un ensemble borné fermé). La fonction $f''(x)$ est uniformément continue sur Q, donc $\| f''_{k-i} - f''_k \| = \lambda_{k-i} \to 0$ et $\sup\limits_{0 \leqslant \tau \leqslant 1} \| f''(x_{k-i} + \tau r_{k-i}) - f''(x_{k-i}) \| = \mu_{k-i} \to 0$ pour $k \to \infty$.

Ainsi, il découle de (3.7)

$$\| B_k r_{k-i} \| \leqslant (\lambda_{k-i} + \mu_{k-i}) \| r_{k-i} \| = h_{k-i} \| r_{k-i} \|, \tag{3.8}$$

où $h_{k-i} \to 0$ quand $k \to \infty$.

On a par définition de la norme d'un opérateur: $\| B_k \| = \max\limits_{\|z\|=1} \| B_k z \|$.

Soit z_k l'élément réalisant le maximum. Si

$$z_k = \delta_k \frac{r_k}{\| r_k \|} + \ldots + \delta_{k-n+1} \frac{r_{k-n+1}}{\| r_{k-n+1} \|}$$

en vertu de la condition $|\Delta_k| \geqslant \varepsilon > 0$ les coefficients δ_{k-i} sont bornés: $|\delta_{k-i}| \leqslant C$, $i = 0, 1, \ldots, n-1$. Moyennant la représentation du vecteur z_k on obtient

$$\| B_k \| = \| B_k z_k \| = \left\| \sum_{i=0}^{n-1} \delta_{k-i} B_k \frac{r_{k-i}}{\| r_{k-i} \|} \right\| \leqslant \sum_{i=0}^{n-1} \left\| \delta_{k-i} B_k \frac{r_{k-i}}{\| r_{k-i} \|} \right\|$$

D'où, compte tenu de (3.8) et de la borne de $|\delta_{k-i}|$,

$$\| B_k \| \leqslant \sum_{i=0}^{n-1} |\delta_{k-i}| \frac{h_{k-i} \| r_{k-i} \|}{\| r_{k-i} \|} = \sum_{i=0}^{n-1} |\delta_{k-i}| h_{k-i} \to 0$$

pour $k \to \infty$. Le lemme est démontré.

Le lemme ouvre en fait la voie pour la construction des méthodes du type (3.4).

LEMME 3.2. *Si $f(x)$ est une fonction fortement convexe continûment dérivable et la suite $\{x_k\}$ est telle que $f_{k+1} \leqslant f_k$ et $(f'_k, x_{k+1} - x_k) \to 0$ pour $k \to \infty$, alors $\| x_{k+1} - x_k \| \to 0$.*

DÉMONSTRATION. Par suite de la condition $f_{k+1} \leqslant f_k$, on a, quel que soit k, $x_{k+1} \in S_k$, $S_k = \{x: f(x) \leqslant f(x_k)\}$. L'ensemble S_k est fortement convexe vu la convexité forte de $f(x)$ (lemme I.2.8). Il existe donc un nombre positif $\lambda > 0$ tel que tout point $\frac{x_{k+1}+x_k}{2} + \xi$, où $\|\xi\| \leqslant \lambda \|x_{k+1} - x_k\|^2$, soit intérieur à S_k. Soit $\frac{x_{k+1}-x_k}{2} = v + \omega$, où $v \in T_k$, T_k étant un plan tangent à S_k en x_k, et $\omega \perp T_k$. Du moment que $f'(x_k) \perp T_k$, on obtient dans ce cas

$$\frac{1}{2} |(f'_k, x_{k+1} - x_k)| = |(f'_k, v + \omega)| = \|f'_k\| \|\omega\|.$$

Mais $\|\omega\| > \|\xi\|$, sinon l'ensemble S_k et le plan T_k posséderaient en commun des points en plus de x_k, ce qui contredit la convexité stricte de S_k. Aussi

$$\frac{1}{2} |(f'_k, x_{k+1} - x_k)| \geqslant \lambda \|f'_k\| \|x_{k+1} - x_k\|^2.$$

Si $\|f'_k\| \not\to 0$, il en résulte $\|x_{k+1} - x_k\| \to 0$. Si, par contre, $\|f'_k\|$ tend vers zéro, la forte convexité de $f(x)$ entraîne $d_k \to 0$, d_k étant le diamètre maximal de S_k, et donc $\|x_{k+1} - x_k\| \to 0$. Le lemme est démontré.

Examinons maintenant de près les propriétés du processus (3.4). Nous l'étudierons en supposant la valeur du paramètre α_k choisie à partir de la condition (2.2) (§ 2) et en nous rappelant que dans ce cas $p_k = -A_k^{-1} f'_k$. La fonction à minimiser est supposée régulière fortement convexe. Le choix de α_k est justifié d'après le même schéma qu'aux paragraphes précédents. Le développement taylorien nous donne l'estimation

$$f_{k+1} - f_k \leqslant \alpha_k (f'_k, p_k) \left[1 + \frac{\alpha_k}{2} \frac{(A_k p_k, p_k)}{(f'_k, p_k)} + \frac{\alpha_k a_k}{2} \frac{\|p_k\|^2}{(f'_k, p_k)} \right],$$

où $a_k = \|f''_{kc} - A_k\|$, $x_{kc} = x_k + \theta (x_{k+1} - x_k)$, $\theta \in [0, 1]$. Notons que

$$(f'_k, p_k) = -(A_k p_k, p_k) = -(f'_k, A_k^{-1} f'_k). \tag{3.9}$$

Vu (3.9),

$$f_{k+1} - f_k \leqslant \alpha_k (f'_k, p_k) \left[1 - \frac{\alpha_k}{2} + \frac{\alpha_k a_k}{2} \frac{\|p_k\|^2}{(f'_k, p_k)} \right],$$

qui implique que la condition (2.2) est manifestement remplie si α_k vérifie

$$1 - \frac{\alpha_k}{2} + \frac{\alpha_k a_k}{2} \frac{\|p_k\|^2}{(f'_k, p_k)} \geqslant \varepsilon. \tag{3.10}$$

THÉORÈME 3.1. *Si $f(x)$ est une fonction deux fois continûment dérivable vérifiant les conditions* (2.4), *la matrice A_k se définit pour*

tout $k \geqslant n-1$ *par le système* (3.6) *et satisfait à*

$$(A_k^{-1} f_k', f_k') > 0 \tag{3.11}$$

et α_k *est défini par la condition* (2.2), *alors, quel que soit le point initial* x_0, *la suite* (3.4) *est telle que* $f_{k+1} < f_k$ *et* $\| x_k - x_* \| \to 0$, *et la convergence est superlinéaire*:

$$\| x_{N+l} - x_* \| \leqslant C\lambda_N \ldots \lambda_{N+l}. \tag{3.12}$$

Ici C, $N < \infty$, $\lambda_{N+l} < 1$ *pour tout* $l \geqslant 0$, $\lambda_i \to 0$ *lorsque* $i \to \infty$.

Démonstration. On ne saurait utiliser le lemme 3.1 qu'après avoir montré que sous les hypothèses du théorème la suite (3.4) vérifie $\| x_{k+1} - x_k \| \to 0$.

En vertu des conditions (3.9) et (3.11), on a $(f_k', p_k) < 0$ pour k quelconque. Il en découle, primo, qu'il existe toujours une valeur $\alpha_k \neq 0$ pour laquelle on a l'inégalité (3.10) (et donc (2.2)) et, secundo, que $f_{k+1} < f_k$ par suite de (2.2). Cela signifie que $x_{k+1} \in S = \{x : f(x) \leqslant f(x_0)\}$ pour tout k et que, $f(x)$ étant minorée, $f_k - f_{k+1} \to 0$. Par suite, il résulte de (2.2)

$$\alpha_k (f_k', p_k) = (f_k', x_{k+1} - x_k) \to 0. \tag{3.13}$$

Puisque $f_{k+1} \leqslant f_k$ et (3.13) a lieu, la suite $\{x_k\}$ satisfait aux exigences du lemme 3.2. Par conséquent, pour $k \to \infty$,

$$\| x_{k+1} - x_k \| \to 0. \tag{3.14}$$

Ainsi, les conditions du théorème garantissent toutes les exigences du lemme 3.1, et donc, pour $k \to \infty$,

$$\| A_k - f_k'' \| \to 0. \tag{3.15}$$

Compte tenu de (2.4), il s'ensuit que, quels que soient M_1 et m_1 tels que $M_1 \geqslant M$ et $0 < m_1 \leqslant m$, on trouve un nombre L tel que, quand $k \geqslant L$, on ait pour tout $y \in E^n$ (1)

$$m_1 \| y \|^2 \leqslant (A_k y, y) \leqslant M_1 \| y \|^2. \tag{3.16}$$

En vertu de (3.9) et (3.16), à partir d'un certain k on a $(f_k', p_k) \leqslant -m_1 \| p_k \|^2$. L'inégalité (3.10) a donc manifestement lieu si l'on a

$$1 - \frac{\alpha_k}{2} - \frac{\alpha_k a_k}{2} \frac{1}{m_1} \geqslant \varepsilon, \quad 0 < \varepsilon < \frac{1}{2}. \tag{3.17}$$

Vu que $a_k \leqslant M + M_1 < \infty$, on se convainc facilement qu'il existe une constante $\bar{\alpha} > 0$ telle que, quel que soit k, l'inégalité (3.17) soit satisfaite pour $\alpha_k \geqslant \bar{\alpha}$. Par suite, (3.14) entraîne $\| p_k \| = \frac{1}{\alpha_k} \| x_{k+1} - x_k \| \to 0$. D'où

$$\| f_k' \| = \| A_k p_k \| \leqslant M_1 \| p_k \| \to 0.$$

Conformément à (1.12), cette dernière condition équivaut à $x_k \to x_*$.

Etablissons l'estimation (3.12).

Etant donné que $\| p_k \| \to 0$, qu'on a (3.15) et que les dérivées secondes de $f(x)$ sont uniformément continues sur S, on a pour $k \to \infty$

$$a_k \leqslant \| f''(x_k + \theta(x_{k+1} - x_k)) - f''(x_k) \| + \| f''(x_k) - A_k \| \to 0,$$

qui implique que, quel que soit $0 < \varepsilon < 1/2$, l'inégalité (3.17) est valable pour $\alpha_k = 1$ à partir d'une certaine itération. Le processus (3.4) se réalise donc avec le pas unité.

La première estimation (3.16) signifie qu'on a $\| A_k^{-1} \| \leqslant \frac{1}{m_1}$ pour $k \geqslant L$ (lemme I.2.9). Jointe à l'invertibilité de A_k cette estimation permet de conclure pour tout $k \geqslant n - 1$ qu'il existe une constante M_2 telle que $\| A_k^{-1} \| \leqslant M_2$ pour $k \geqslant n - 1$. Fort de ce résultat on peut évaluer la vitesse de convergence comme pour le théorème 2.1 :

$$\| x_{k+1} - x_* \| \leqslant \| I - A_k^{-1} f''_{kc} \| \, \| x_k - x_* \| \leqslant \| A_k^{-1} \| \times$$
$$\times \| A_k - f''_{kc} \| \, \| x_k - x_* \| \leqslant M_2 \| A_k - f''_{kc} \| \, \| x_k - x_* \|,$$

ou $\| x_{k+1} - x_* \| \leqslant \lambda_k \| x_k - x_* \|$, où $\lambda_k = M_2 \| A_k - f''_{kc} \|$. Puisque $\| A_k - f''_{kc} \| \leqslant \| A_k - f''_k \| + \| f''_k - f''(x_k + \theta(x_k - x_*)) \| \to 0$, on trouve un nombre N tel qu'on ait $\lambda_{N+l} < 1$ pour $k = N + l$, $l = 0, 1, \ldots$, et $\lambda_{N+l} \to 0$ avec l tendant vers l'infini.

En posant $\| x_N - x_* \| = C$ et en profitant de ce que nous avons dit sur les valeurs λ_k, nous arrivons à l'estimation (3.12), c.q.f.d.

La condition (3.11) du théorème signifie, en vertu de (3.9), que p_k est une direction de descente de $f(x)$. Il se peut que $(A_k^{-1} f'_k, f'_k) \leqslant 0$ à certaines itérations. On modifie alors le vecteur r_k et on reconstruit la matrice A_k^{-1} (de façon à remplir la condition (3.11)) ou on fait un pas dans la direction de l'opposé du gradient. Les pas seront toujours en nombre fini parce qu'en suivant cette direction on a $\| x_{k+1} - x_k \| \to 0$ et $\| A_k - f''_k \| \to 0$ (si le vecteur r_k vérifie les conditions formulées plus haut). En vertu de (3.16) et (3.9) on a donc nécessairement, à partir d'une certaine itération, $(A_k^{-1} f'_k, f'_k) > 0$. Il serait d'ailleurs plus simple (si $(A_k^{-1} f'_k, f'_k) = -(p_k, f'_k) < 0$ pour un k) de faire changer de signe le facteur scalaire α_k : on se déplacera alors à partir de x_k dans la direction $-p_k$, i.e. dans une direction de descente.

3. Construction de divers algorithmes. Les conditions auxquelles sont assujettis r_k servant à générer la suite (3.5) ne sont pas sévères et elles laissent un grand arbitraire sur le choix de ces vecteurs. Cela permet de construire des algorithmes du type (3.4) puisque les différentes suites $\{r_k\}$ définissent (selon (3.6)) autant de suites de matrices A_k.

Voyons certains procédés de formation des r_k. On peut prendre pour ceux-ci des vecteurs dirigés suivant les axes de coordonnées. Si, par exemple, $r_0 = \lambda_0 v_1$, alors, pour $k = tn + i$, t entier et $i = 0, 1, \ldots, n-1$, on a $r_k = \lambda_k v_{i+1}$. Ici v_{i+1} est le vecteur unité de l'axe correspondant et λ_k, un facteur numérique qui doit remplir la condition : $\lambda_k \to 0$ pour $k \to \infty$. Ce choix des r_k garantit la condition $|\Delta_k| \geqslant \varepsilon$. Dans ce cas la définition de la matrice A_k demande à chaque itération le calcul des dérivées en deux points x_k et y_k. La loi de décroissance de λ_k peut être quelconque, mais la pratique montre que la convergence la plus rapide a lieu pour sa décroissance monotone ; on peut poser par exemple $\lambda_k = 1/k^2$.

Un autre procédé de définition des r_k consiste à utiliser, pour $k \geqslant n-1$, non pas (3.5) mais la suite (3.4), i.e. on pose $r_k = x_{k+1} - x_k = -\alpha_k A_k^{-1} f'_k$.

En effet, la démonstration du théorème 3.1 montre qu'avec A_k une matrice arbitraire vérifiant la seule condition (3.11) et α_k choisi à partir de la condition (2.2), on a $\| x_{k+1} - x_k \| \to 0$ pour $k \to \infty$. En générant les vecteurs r_k moyennant la suite (3.4) on a donc forcément la condition $\| r_k \| \to 0$ et il ne reste qu'à tester la condition $|\Delta_k| \geqslant \varepsilon$. Si elle n'a pas lieu pour un k, on a à choisir un autre r_k (cette fois-ci non pas d'après (3.4)). Dans un tel algorithme, pour définir la matrice A_k à chaque itération (où la suite (3.4) garantit les conditions imposées aux vecteurs r_k), il faut calculer le gradient au seul point x_k.

Les deux procédés décrits ne sont certes pas uniques.

Le système d'équations (3.6) donnant A_k n'utilise pour tout k qu'un seul nouveau r_k et e_k correspondant, les autres $r_{k-1}, \ldots, r_{k-n+1}$ et $e_{k-1}, \ldots, e_{k-n+1}$ étant formés aux itérations précédentes. On modifie le système (3.6) de façon qu'à chaque itération du processus (3.4) il apparaît dans (3.6) un nombre arbitraire de vecteurs $r_{k-i_1}, \ldots, r_{k-i_j}$, $1 \leqslant j \leqslant n$, (et de vecteurs $e_{k-i_1}, \ldots, e_{k-i_j}$ correspondants) nouveaux, les $n-j$ autres vecteurs $r_{k-i_{j+1}}, \ldots, r_{k-i_n}$ étant construits dès itérations précédentes. On met alors le système (3.6) sous une forme plus commode

$$A_k r_i = e_i, \quad i = 1, \ldots, n. \tag{3.18}$$

Si les exigences du lemme 3.1 sont maintenues et si on reprend exactement sa démonstration, on s'assure que la matrice A_k définie par le système (3.18) vérifie aussi la condition $\| A_k - f''_k \| \to 0$ pour $k \to \infty$.

En appliquant dans (3.18) divers procédés de construction des r_i on aboutit à des algorithmes de minimisation classiques. Ainsi, en posant $r_i = v_{ki}$ (auquel cas $y_i = x_k + v_{ki}$), où v_{ki} est un vecteur dirigé suivant le i-ième axe de coordonnées et tel que $\| v_{ki} \| \to 0$

pour $k \to \infty$, le système (3.18) se récrit

$$A_k v_{ki} = f'(x_k + v_{ki}) - f'(x_k), \quad i = 1, \ldots, n.$$

La matrice A_k ainsi définie est l'analogue aux différences de la matrice des dérivées secondes $f''(x_k)$ et le processus (3.4) devient donc l'analogue aux différences de la méthode de Newton à pas variable. En se rappelant le théorème 3.1 on dit que cet analogue converge à partir d'une approximation initiale quelconque avec une vitesse superlinéaire. Sous l'hypothèse que $f''(x)$ vérifie la condition de Lipschitz (2.8) on montre à l'aide des résultats précédents que si $\| v_{ki} \| \leqslant \| f_k'' \|$, la convergence est quadratique. En effet, dans le lemme 3.1 l'estimation de la quantité $B_k r_i = B_k v_{ki}$ prend la forme

$$\| B_k v_{ki} \| \leqslant \sup_{0 \leqslant \tau \leqslant 1} \| f''(x_k + \tau v_{ki}) - f''(x_k) \| \, \| v_{ki} \|,$$

d'où, compte tenu de (2.8), $\| B_k v_{ki} \| \leqslant R \| v_{ki} \|^2$. Utilisons cette inégalité et l'estimation $\| v_{ki} \| \leqslant \| f_k' \|$ comme dans le lemme 3.1, nous établissons

$$\| A_k - f_k'' \| = \| B_k \| \leqslant R \sum_{i=1}^{n} | \delta_i | \, \| v_{ki} \| \leqslant R \sum_{i=1}^{n} | \delta_i | \, \| f_k' \| = \widetilde{R} \| f_k' \|.$$

En vertu de (2.4), le gradient $f'(x)$ satisfait à la condition de Lipschitz de rapport M. Donc

$$\| B_k \| \leqslant \widetilde{R} \| f_k' \| = \widetilde{R} \| f_k' - f_*' \| \leqslant \widetilde{R} M \| x_k - x_* \|.$$

Maintenant on améliore comme suit l'estimation de la vitesse de convergence obtenue dans le théorème 3.1 :

$$\begin{aligned} \| x_{k+1} - x_* \| &\leqslant M_2 \| A_k - f_{kc}'' \| \, \| x_k - x_* \| \leqslant \\ &\leqslant M_2 (\| A_k - f_k'' \| + \| f_k'' - f''(x_k + \theta (x_k - x_*)) \|) \times \\ &\qquad \times \| x_k - x_* \| \leqslant M_2 (\widetilde{R} M + R) \| x_k - x_* \|^2, \end{aligned}$$

i.e.

$$\| x_{k+1} - x_* \| \leqslant C \| x_k - x_* \|^2.$$

Donnons un autre procédé de choix des vecteurs r_i. Soit $r_i = -\lambda_k f'(y_i)$, $y_1 = x_k$, $y_{i+1} = y_i + r_i$, $i = 1, 2, \ldots, n$. Le système (3.18) prend alors la forme

$$-\lambda_k A_k f'(y_i) = f'(y_{i+1}) - f'(y_i), \quad i = 1, \ldots, n.$$

La matrice A_k ainsi définie (à un facteur numérique λ_k près) sert à construire le processus itératif d'Aitken-Steffensen. Nous ne nous occuperons pas des propriétés de cette méthode. Bornons-nous à noter qu'en réglant la valeur de λ_k on arrive à une convergence quadratique.

4. Définition du vecteur p_k. Le volume de calcul nécessaire pour trouver p_k détermine essentiellement le coût du processus (3.4).

Nous allons construire le vecteur $p_k = -A_k^{-1}f_k'$ en nous servant des particularités du système (3.6) donnant A_k. Nous réduirons ainsi considérablement le coût de cette construction. Commençons par l'inversion de la matrice A_k.

Une condition nécessaire d'existence de A_k^{-1} est la non-dégénérescence de A_k, ce qui exige à son tour l'indépendance linéaire du système de vecteurs $e_k, \ldots, e_{k-n+1}$. Pour k suffisamment grand la matrice A_k est régulière selon les conditions (3.16). Or, il se peut qu'à certaines premières itérations de (3.4) le système de vecteurs $e_k, \ldots, e_{k-n+1}$ soit linéairement dépendant. On procède alors comme suit: ou bien on modifie un des r_{k-i} ou bien on fait un pas suivant l'opposé du gradient, ce qui modifie le système en question. Nous supposerons dans la suite l'indépendance linéaire du système $e_k, \ldots, e_{k-n+1}$ pour tout $k \geqslant n-1$.

Le système (3.6) se récrit alors

$$A_k^{-1}e_{k-i} = r_{k-i}, \quad i = 0, 1, \ldots, n-1.$$

ou, sous forme d'équation matricielle,

$$A_k^{-1}E_k = R_k, \tag{3.19}$$

où E_k, R_k sont des matrices dont les colonnes sont formées par les coordonnées des vecteurs e_{k-i} et r_{k-i} respectivement.

On obtient à partir de (3.19):

$$A_k^{-1} = R_k E_k^{-1}. \tag{3.20}$$

La construction de A_k^{-1} exige donc en premier lieu qu'on calcule E_k^{-1}. On trouve dans les cours d'algèbre linéaire (voir, p.ex., D. Faddeev et V. Faddeeva [57]) que les lignes de la dernière matrice sont les vecteurs de la base $s_k, \ldots, s_{k-n+1}$ duale (ou biorthogonale) de la base $e_k, \ldots, e_{k-n+1}$. Rappelons que deux systèmes linéairement indépendants de vecteurs $a_1, \ldots, a_n$ et $b_1, \ldots, b_n$ sont dits *duals* (ou biorthogonaux) sous les conditions

$$(a_i, b_j) = 0 \quad \text{pour} \quad i \neq j, \quad (a_i, b_i) = 1.$$

Si la base $s_k, \ldots, s_{k-n+1}$ est duale de $e_k, \ldots, e_{k-n+1}$, alors, en vertu des relations de dualité, $S_k^* E_k = I$, avec S_k une matrice de colonnes composées de vecteurs s_{k-i}. D'où $S_k^* = E_k^{-1}$.

Chacune des E_k, $k = 0, 1, \ldots$, ne diffère de ses deux voisins immédiats que par une colonne. La génération de la base $s_k, \ldots, s_{k-n+1}$ se fait donc moyennant des relations de récurrence, ce qui en réduit sensiblement le coût.

Supposons E_k^{-1} déjà trouvé, i.e. nous avons construit la base $s_k, \ldots, s_{k-n+1}$. Formons le système de vecteurs $\bar{s}_{k+1}, \bar{s}_k, \ldots, \bar{s}_{k-n+2}$:

$$\bar{s}_{k+1} = \frac{s_{k-n+1}}{(s_{k-n+1}, e_{k+1})}, \quad \bar{s}_{k+1-j} = s_{k+1-j} - (s_{k+1-j}, e_{k+1})\,\bar{s}_{k+1},$$
$$j = 1, \ldots, n-1. \tag{3.21}$$

L'indépendance linéaire des $e_{k+1}, e_k, \ldots, e_{k-n+2}$ entraîne

$$(s_{k-n+1}, e_{k+1}) \neq 0. \tag{3.22}$$

En effet, vu la dualité des bases $s_k, \ldots, s_{k-n+1}$ et $e_k, \ldots, e_{k-n+1}$ on a $(s_{k-n+1}, e_{k-j}) = 0$ pour $j = 0, 1, \ldots, n-2$, et si $(s_{k-n+1}, e_{k+1}) = 0$, il en découlerait la dépendance linéaire des $e_{k+1}, \ldots, e_{k-n+2}$. Par conséquent, *pour vérifier l'indépendance linéaire des vecteurs* $e_{k+1}, e_k, \ldots, e_{k-n+2}$, *il suffit de tester la condition* (3.22).

Assurons-nous que le système (3.21) constitue la base duale de $e_{k+1}, \ldots, e_{k-n+2}$.

En effet,

$$(\bar{s}_{k+1}, e_{k+1}) = \frac{(s_{k+1-n}, e_{k+1})}{(s_{k+1-n}, e_{k+1})} = 1,$$

$$(\bar{s}_{k+1-j}, e_{k+1}) = (s_{k+1-j}, e_{k+1}) - (s_{k+1-j}, e_{k+1}) \frac{(s_{k+1-n}, e_{k+1})}{(s_{k+1-n}, e_{k+1})} = 0,$$

$$(\bar{s}_{k+1}, e_{k+1-j}) = \frac{(s_{k+1-n}, e_{k+1-j})}{(s_{k+1-n}, e_{k+1})} = 0$$

vu la dualité des bases $e_k, \ldots, e_{k-n+1}$ et $s_k, \ldots, s_{k-n+1}$,

$$(\bar{s}_{k+1-j}, e_{k+1-m}) = (s_{k+1-j}, e_{k+1-m}) - (s_{k+1-j}, e_{k+1}) \frac{(s_{k+1-n}, e_{k+1-m})}{(s_{k+1-n}, e_{k+1})} = \delta_{jm},$$

$j, m = 1, 2, \ldots, n-1$, δ_{jm} est le symbole de Kronecker ($\delta_{jj} = 1$, $\delta_{jm} = 0$ pour $j \neq m$). Par suite, les systèmes de vecteurs $e_{k+1}, \ldots, e_{k-n+2}$ et $\bar{s}_{k+1}, \ldots, \bar{s}_{k-n+2}$ remplissent les conditions de dualité, d'où notre affirmation.

Ainsi, la construction de la base $\bar{s}_{k+1}, \ldots, \bar{s}_{k-n+2}$ (i.e. de la matrice E_{k+1}^{-1}) moyennant les relations récurrentes (3.21) s'effectue fort simplement.

Il est maintenant possible d'obtenir une formule élémentaire donnant la direction de déplacement p_k. A cette fin mettons l'équation (3.20) sous la forme

$$A_k^{-1} = \sum_{i=0}^{n-1} r_{k-i} s_{k-i}^* \tag{3.23}$$

(l'indice * désigne un vecteur-ligne).

Utilisons cette expression, il vient

$$p_k = -A_k^{-1} f_k' = -\sum_{i=0}^{n-1} r_{k-i} s_{k-i}^* f_k' = -\sum_{i=0}^{n-1} (s_{k-i}, f_k') r_{k-i}. \tag{3.24}$$

C'est précisément cette formule de la direction de descente dans les méthodes du type (3.4) qui est à la base de l'appellation « méthodes des directions duales ». Compte tenu de (3.24), la formule (3.4) peut s'écrire

$$x_{k+1} = x_k - \alpha_k \sum_{i=0}^{n-1} (s_{k-i}, f_k') r_{k-i} \tag{3.25}$$

ou, en coordonnées,

$$x_{k+1}^{\nu} = x_k^{\nu} - \alpha_k \sum_{i=0}^{n-1} \sum_{j=1}^{n} s_{k-i}^{j} \frac{\partial f(x_k)}{\partial x^j} r_{k-i}^{\nu},$$
$$\nu = 1, \ldots, n.$$

En fait, c'est cette formule-là qu'il faut utiliser pour obtenir des approximations successives.

Notons de plus que l'expression (3.23) permet d'obtenir la formule récurrentielle de calcul de A_k^{-1}. Nous la citons toute prête :

$$A_{k+1}^{-1} = A_k^{-1} + (r_{k+1} - A_k^{-1} e_{k+1}) \bar{s}_{k+1}^{*}.$$

Sa validité s'établit sans peine si on multiplie directement la matrice ainsi construite par les vecteurs $e_{k+1}, e_k, \ldots, e_{k-n+2}$. Il se trouve que

$$A_{k+1}^{-1} e_{k+1-i} = r_{k+1-i}, \quad i = 0, 1, \ldots, n-1,$$

i.e. la matrice A_{k+1}^{-1} satisfait au système (3.6).

5. **Démarrage du processus.** Le processus itératif (3.4) n'a débuté jusqu'à présent qu'avec $k = n - 1$ vu que pour définir A_k il faut n vecteurs r_k et autant de vecteurs e_k associés.

Les premières itérations ($k < n - 1$) s'effectuent de différentes façons. On utilise par exemple la méthode de la plus grande pente : $x_{k+1} = x_k - \alpha_k f'_k$, $\alpha_k > 0$, $k = 0, 1, \ldots, n-2$. L'algorithme se réalisera uniformément à partir de la première itération si, pour $0 \leqslant k < n - 1$, le calcul est organisé comme suit.

Posons $A_0^{-1} = I$ et mettons-la sous la forme $A_0^{-1} = R_0 E_0^{-1}$, où $R_0 = I$, $E_0^{-1} = I$, ou (en recourant à (3.23))

$$A_0^{-1} = \sum_{i=0}^{n-1} r_{0-i} s_{0-i}^{*},$$

où $r_0, r_{-1}, \ldots, r_{-n+1}$ et $s_0, s_{-1}, \ldots, s_{-n+1}$ sont éléments des bases orthonormées unitaires. Par suite,

$$x_1 = x_0 - \alpha_0 \sum_{i=0}^{n-1} (f'_0, s_{0-i}) r_{0-i}.$$

Calculons les vecteurs r_1 et e_1 et construisons, moyennant les formules (3.21), la base

$$\bar{s}_1 = \frac{s_{-n+1}}{(s_{-n+1}, e_1)}, \quad \bar{s}_{1-j} = s_{1-j}, \quad -(s_{1-j}, e_1)\bar{s}_1, \quad j = 1, \ldots, n-1,$$

et l'approximation suivante :

$$x_2 = x_1 - \alpha_1 \sum_{i=0}^{n-1} (f'_1, \bar{s}_{1-i}) r_{1-i}.$$

Il est immédiat de construire les itérations suivantes.

6. Minimisation d'une forme quadratique. Illustrons les méthodes des directions duales pour la recherche du minimum d'une fonction quadratique. Soit

$$f(x) = \frac{1}{2}(Ax, x) + (b, x) + c.$$

Ici A est une $n \times n$-matrice symétrique strictement définie positive à éléments constants : $(Ax, x) > 0$ quel que soit $x \neq 0$; b un vecteur et c un scalaire. Le gradient de cette fonction est $f'(x) = Ax + b$ et le vecteur

$$e_i = f'(x + r_i) - f'(x) = Ar_i. \tag{3.26}$$

Par conséquent, si $r_1, \ldots, r_n$ est un système linéairement indépendant et $s_1, \ldots, s_n$ la base duale de $e_1, \ldots, e_n$, alors on a, en vertu de (3.20) et (3.23),

$$A_n^{-1} = R_n E_n^{-1} = \sum_{i=1}^{n} r_i s_i^*.$$

Puisque, conformément à (3.26), la matrice A se définit par le système d'équations $Ar_i = e_i$, $i = 1, 2, \ldots, n$, on a

$$A^{-1} = R_n E_n^{-1} = \sum_{i=1}^{n} r_i s_i^*, \tag{3.27}$$

i.e. $A_n^{-1} = A^{-1}$. Donc

$$x_{n+1} = x_n - A_n^{-1} f_n' = x_n - A^{-1}(Ax_n + b) = -A^{-1}b \tag{3.28}$$

et $f_{n+1}' = -AA^{-1}b + b = 0$, i.e. $x_{n+1} = x_*$.

Pour minimiser une fonction quadratique par les méthodes des directions duales il faut donc calculer son gradient en $n + 1$ points et construire la duale de la base de vecteurs $e_1, \ldots, e_n$. Si l'on considère le calcul successif des $e_1, \ldots, e_n$ comme une procédure itérative, on peut dire que les méthodes en question permettent de réaliser le minimum en un nombre fini de pas.

Remarquons pour conclure que le problème examiné équivaut à la résolution du système d'équations linéaires $Ax = -b$. Avec les techniques utilisant des directions duales on résout donc un système d'équations linéaires au bout d'un nombre fini d'itérations.

7. Discussion des propriétés des méthodes. Les méthodes des directions duales permettent de trouver le minimum d'une fonction régulière strictement convexe à partir de toute approximation initiale, la convergence de la suite $\{x_k\}$ vers la solution étant superlinéaire. Le procédé de choix du paramètre α_k garantit l'obtention de la valeur cherchée de α_k après n fractionnements. Il est vrai que

dans le processus (3.4) comme dans les méthodes des paragraphes précédents α_k peut être choisi à partir de la condition de minimum dans la direction de déplacement, mais cette façon d'agir est plus laborieuse.

En ce qui concerne l'évaluation de la vitesse de convergence, les méthodes de la classe considérée s'approchent de la méthode de Newton. Comparons leur coût par itération. Dans les processus du type (3.4) avec A_k définie par le système (3.6) la recherche de A_k^{-1} nécessite le calcul du vecteur e_k et la construction, selon les formules récurrentielles (3.21), de la base duale de $e_k, \ldots, e_{k-n+1}$.

On construit le vecteur e_k (n° 3) en calculant le gradient de la fonction en un ou deux points. Du fait de former la base duale moyennant (3.21) on en diminue le coût de n fois (et même de plus si on vérifie les conditions de non-nullité du dénominateur dans les formules générales régissant la construction de la base duale) par rapport à des procédés habituels (D. Faddeev et V. Faddeeva [57]).

On constate donc en premier lieu qu'à la différence de la méthode de Newton, celles des directions duales n'exigent pas qu'on trouve des dérivées secondes de la fonction. Une comparaison avec l'analogue aux différences de la première méthode montre que la construction de la matrice A_k^{-1} coûte environ n fois moins pour les directions duales parce que dans la méthode de Newton on calcule les dérivées en $n + 1$ points par itération et on inverse la matrice A_k sans recourir à des relations de récurrence.

Au lieu d'inverser A_k, on définit la direction de déplacement p_k dans la méthode de Newton aux différences en résolvant un système d'équations linéaires (de même que dans la version classique). Dans ce cas, l'évaluation quantitative des coûts relatifs des processus du type (3.4) et de la méthode de Newton est fonction du nombre d'opérations nécessaires pour résoudre le système d'équations, mais elle est, dans ce cas également, environ n. Si l'on veut par exemple résoudre le système par des directions duales (n° 6), on a en fait à calculer la matrice A_n^{-1} sans utiliser les relations récurrentes, ce qui demande, nous l'avons dit, n fois plus de calculs que lorsqu'on le fait moyennant les formules (3.21).

La résolution du système d'équations linéaires par les méthodes des directions conjuguées (voir paragraphe suivant) exige sensiblement le même effort.

Ainsi, les méthodes des directions duales exigent, à convergence à peu près égale, beaucoup moins de calculs par itération que la méthode de Newton.

Elles présentent cependant l'inconvénient de provoquer, quand il s'agit du calcul sur machine, un grand encombrement de mémoire puisqu'il faut stocker deux systèmes de vecteurs $r_k, r_{k-1}, \ldots, r_{k-n+1}$ et $s_k, s_{k-1}, \ldots, s_{k-n+1}$, i.e. deux matrices $n \times n$. Cela restreint leur emploi dans la résolution de problèmes de plusieurs variables sur

machines à mémoire rapide de capacité limitée. On y pallie en partie en prenant pour r_k des vecteurs dirigés suivant les axes de coordonnées auquel cas on ne conserve en mémoire qu'un seul n-vecteur (au lieu du système $r_k, \ldots, r_{k-n+1}$).

§ 4. Méthodes des directions conjuguées. Minimisation de fonctions quadratiques

1. Directions conjuguées et leurs propriétés. Reprenons le problème de minimisation de la fonction quadratique

$$f(x) = \frac{1}{2}(Ax, x) + (b, x) + c, \tag{4.1}$$

où $(Ax, x) > 0$ pour tout $x \neq 0$ (n° 6, § 3). On constate aisément que la minimisation d'une fonction quadratique se ramène en fait à l'inversion de A : connaissant A^{-1} la solution s'obtient immédiatement par les formules (3.28)

$$x_* = x_0 - A^{-1}f'_0 = -A^{-1}b, \tag{4.2}$$

x_0 étant un point quelconque.

Pour calculer A^{-1} à l'aide des expressions (3.27), il faut, après avoir choisi un système linéairement indépendant $p_0, \ldots, p_{n-1}$ (nous notons ici p_i au lieu de r_i), trouver les vecteurs correspondants

$$\tilde{e}_i = f'(x_i + p_i) - f'(x_i) = Ap_i, \quad i = 0, 1, \ldots, n-1, \tag{4.3}$$

x_i étant des points arbitraires, et construire une base $s_0, \ldots, s_{n-1}$ duale de $\tilde{e}_0, \ldots, \tilde{e}_{n-1}$, i.e. telle que

$$(s_i, \tilde{e}_i) = 1, \quad (s_i, \tilde{e}_j) = 0 \quad \text{pour} \quad i \neq j. \tag{4.4}$$

Ces relations s'écrivent également, en vertu de (4.3),

$$(s_i, Ap_i) = 1, \quad (s_i, Ap_j) = 0, \quad i \neq j. \tag{4.5}$$

Un cas particulièrement intéressant est celui des *vecteurs* $p_0, \ldots, p_{n-1}$ *A-conjugués* (ou *conjugués* tout court), i.e. *vérifiant les conditions*

$$(p_i, Ap_j) = 0, \quad i \neq j. \tag{4.6}$$

Un système de vecteurs $p_0, \ldots, p_{n-1}$ (*non nuls*) *remplissant les conditions* (4.6) *est linéairement indépendant* (en tant que systhème orthogonal dans la métrique définie par une matrice régulière), et il peut donc être utilisé pour définir e_i par les formules (4.3). Les vecteurs s_i vérifiant (4.5) se calculent alors selon les formules fort simples :

$$s_i = \frac{p_i}{(Ap_i, p_i)}, \quad i = 0, 1, \ldots, n-1. \tag{4.7}$$

Si les vecteurs $p_0, \ldots, p_{n-1}$ sont A-conjugués, la matrice A^{-1} s'obtient donc par la formule (voir (3.27))

$$A^{-1} = \sum_{i=0}^{n-1} p_i s_i^* = \sum_{i=0}^{n-1} \frac{p_i p_i^*}{(Ap_i, p_i)}, \tag{4.8}$$

i.e. le problème d'inversion de A et, par là même, celui de recherche du minimum de $f(x)$ se résout de façon très simple.

Voyons le problème de recherche de x_* par des vecteurs conjugués d'un point de vue légèrement différent et établissons en passant plusieurs propriétés intéressantes des directions conjuguées.

Puisque $p_0, \ldots, p_{n-1}$ constituent une base dans E^n, on représente x_* comme

$$x_* = x_0 + \sum_{i=0}^{n-1} \alpha_i p_i. \tag{4.9}$$

Mais en vertu de (4.2) et (4.8)

$$x_* = x_0 - \sum_{i=0}^{n-1} \frac{p_i p_i^*}{(Ap_i, p_i)} f_0'. \tag{4.10}$$

Il résulte de (4.9) et (4.10)

$$x_0 + \sum_i \alpha_i p_i = x_0 - \sum_i \frac{p_i p_i^*}{(Ap_i, p_i)} f_0'$$

ou, sous une autre forme,

$$x_0 + \sum_i \alpha_i p_i = x_0 - \sum_i \frac{(f_0', p_i)}{(Ap_i, p_i)} p_i. \tag{4.11}$$

Vu l'unicité du développement d'un vecteur suivant la base, la dernière égalité permet de définir les valeurs des coefficients α_k dans le développement (4.9)

$$\alpha_i = -\frac{(f_0', p_i)}{(Ap_i, p_i)} = -\frac{(f_0', p_i)}{(\tilde{e}_i, p_i)}, \quad i = 0, 1, \ldots, n-1. \tag{4.12}$$

Si l'on connaît donc un système de vecteurs conjugués, le point de minimum de la fonction quadratique (4.1) se définit facilement par les formules (4.9), (4.12).

La définition du point x_* par la formule (4.9) peut s'interpréter comme génération successive de points :

$$x_{i+1} = x_i + \alpha_i p_i, \quad i = 0, 1, \ldots, n-1, \tag{4.13}$$

avec α_i donnés par (4.12). Par suite, *le problème de minimisation d'une fonction quadratique se résout par directions conjuguées en un*

nombre fini de pas égal au plus à n (le nombre de points du processus itératif (4.13) peut être inférieur à n si certains α_i du développement (4.9) sont nuls, i.e. si $(f'_0, p_i) = 0$ pour certains i). C'est la propriété la plus importante des directions conjuguées. Elle révèle l'efficacité des vecteurs conjugués dans la recherche du minimum d'une fonction quadratique et explique la grande popularité des méthodes du même nom.

On montre en tant que conséquence intéressante du résultat obtenu que *le point x_i généré par les formules* (4.13), (4.12) *réalise le minimum de la fonction* (4.1) *sur un sous-espace engendré par les vecteurs $p_0, \ldots, p_{i-1}$ et passant par x_0.*

En effet, soit

$$\tilde{x}_i = x_0 + \sum_{k=0}^{i-1} \tilde{\alpha}_k p_k,$$

$\tilde{\alpha}_k$ étant quelconques. Pour que $\tilde{x}_i$ soit point de minimum d'une fonction dérivable strictement convexe sur un sous-espace formé par les vecteurs $p_0, \ldots, p_{i-1}$, il faut et il suffit (conséquence I.4.4) qu'on ait

$$(f'(\tilde{x}_i), p_j) = 0, \quad j = 0, 1, \ldots, i-1. \tag{4.14}$$

Mais quels que soient $0 \leqslant j \leqslant i-1$

$$(f'(\tilde{x}_i)\, p_j) = (A\tilde{x}_i + b, p_j) = \left(A\left(x_0 + \sum_{k=0}^{i-1} \tilde{\alpha}_k p_k\right) + b, p_j\right) =$$

$$= (Ax_0 + b, p_j) + \sum_{k=0}^{i-1} \tilde{\alpha}_k (Ap_k, p_j) = (f'_0, p_j) + \tilde{\alpha}_j (Ap_j, p_j).$$

Il en découle, compte tenu de (4.14), que le point $\tilde{x}_i$ réalise le minimum de la fonction sur un sous-espace formé par $p_0, \ldots, p_{i-1}$ et passant par le point x_0 si (et seulement si) $(f'_0, p_j) + \tilde{\alpha}_j (Ap_j, p_j) = 0$, i.e.

$$\tilde{\alpha}_j = -\frac{(f'_0, p_j)}{(Ap_j, p_j)}.$$

Mais ces coefficients coïncident avec α_i calculés par la formule (4.12), i.e. le point $\tilde{x}_i$ en lequel est atteint le minimum cherché se confond avec x_i (4.13). Donc

$$(f'(x_i), p_j) = 0, \quad j = 0, 1, \ldots, i-1. \tag{4.15}$$

On voit donc que la recherche du minimum d'une fonction quadratique dans l'espace E^n par les formules (4.13), (4.12) peut s'interpréter comme minimisation successive dans des sous-espaces de dimension $i+1$, $i = 0, 1, \ldots, n-1$, chaque point de minimum étant trouvé par le calcul d'un seul coefficient α_i.

Pour obtenir ces derniers moyennant les formules (4.12) on calcule en fait, au lieu de la matrice des dérivées secondes A, les vecteurs $\tilde{e}_i = f'(x_i + p_i) - f'(x_i)$ (voir (4.3)), i.e. les seules dérivées premières de la fonction.

On s'assure sans peine que les formules (4.12) donnant les coefficients α_i sont susceptibles de prendre une autre forme. En effet, définissant x_i par (4.13), on a

$$(f'_0, p_i) = (f'_0 - f'_1 + f'_1 - \ldots - f'_i + f'_i, p_i) =$$
$$= (-\alpha_0 A p_0 - \alpha_1 A p_1 - \ldots - \alpha_{i-1} A p_{i-1} + f'_i, p_i),$$

et, les vecteurs $p_0, \ldots, p_i$ étant A-conjugués, il vient $(f'_0, p_i) = (f'_i, p_i)$. Par suite

$$\alpha_i = -\frac{(f'_i, p_i)}{(Ap_i, p_i)} = -\frac{(f'_i, p_i)}{(\tilde{e}_i, p_i)}, \quad i = 0, 1, \ldots, n-1. \quad (4.16)$$

Il s'ensuit que si pour un certain $0 \leqslant i \leqslant n-1$ on a $\alpha_i = 0$ (i.e. $x_{i+1} = x_i$) dans (4.13), alors $(f'_i, p_i) = 0$. Réunissons cette égalité avec (4.15):

$$(f_{i+1}, p_j) = (f_i, p_j) = 0, \quad j = 0, 1, \ldots, i.$$

Ainsi, l'annulation du coefficient α_i signifie que le point x_i correspondant réalise le minimum de la fonction quadratique sur un sous-espace engendré par les vecteurs $p_0, \ldots, p_i$ et passant par le point x_0.

Enfin, en vertu de (4.15), $(f'_i, p_{i-1}) = 0$, en d'autres termes, le choix des α_i selon les formules (4.12) ou (4.16) équivaut à leur choix à partir de la condition

$$f(x_i + \alpha_i p_i) = \min_{\alpha} f(x_i + \alpha p_i).$$

2. Construction des méthodes. Nous avons parlé plus haut (n°1) de l'efficacité des directions conjuguées lorsqu'il s'agit de minimiser une fonction quadratique, mais nous avons passé sous silence la construction de tels vecteurs et le coût de cette procédure.

Chacun des procédés de construction de vecteurs A-conjugués que nous allons étudier déterminera une méthode des directions conjuguées consistant à former des approximations successives du minimum de la fonction (4.1) à l'aide des formules (4.13), (4.12) (ou (4.16)).

L'efficacité des techniques des directions conjuguées dépendra directement de l'effort de calcul nécessaire pour construire le système de vecteurs conjugués. Si la dernière opération s'avère trop laborieuse, la méthode ainsi définie risque d'être peu puissante (par rapport à des algorithmes d'autres classes). On ferait donc bien de formuler les conditions générales que tout procédé de construction doit vérifier pour avoir une méthode des directions conjuguées suffisamment efficace.

Premièrement, en construisant les vecteurs on peut se borner à calculer la fonction et son gradient et négliger les dérivées secondes. Au cas où cette condition n'est pas satisfaite, la recherche du point de minimum d'une fonction quadratique par la méthode (4.13) peut exiger qu'on calcule la matrice des dérivées secondes et, de plus, les valeurs des gradients en plusieurs points. Aussi une méthode des directions conjuguées impliquant le calcul de la matrice des dérivées secondes est-elle en général moins efficace que celle de Newton (à l'exception éventuelle de problèmes où l'inversion de la matrice A est sensiblement plus laborieuse que son calcul).

Deuxièmement, la construction de vecteurs conjugués se fait avec la seule information sur la fonction aux points de la suite (4.13). Autrement dit, ce processus doit marcher de sorte que le vecteur p_i, $0 \leqslant i \leqslant n-1$, se définisse moyennant la seule information sur les valeurs de la fonction et de son gradient aux points $x_0, \ldots, x_i$.

Cette exigence fait se limiter à des techniques de construction telles que la condition

$$(f'_i, p_i) = 0, \quad 0 \leqslant i \leqslant n-1, \tag{4.17}$$

n'ait lieu que si $f'_i = 0$. En effet, si on a (4.17), alors $\alpha_i = 0$ en vertu de (4.16) et, partant, $x_{i+1} = x_i$ dans la suite (4.13). Cela signifie que la $(i+1)$-ième itération ne fournira aucune information nouvelle sur la fonction et qu'il sera donc impossible à ce stade de construire un vecteur $p_{i+1} \neq p_i$. Le processus s'arrêtera donc avant d'atteindre la solution si $f'_i \neq 0$.

Ainsi, tout procédé de construction de vecteurs conjugués (et la méthode associée des directions conjuguées) doit vérifier la condition

$$(f'_i, p_i) \neq 0 \quad \text{si} \quad f'_i \neq 0. \tag{4.18}$$

Cette condition garantit $\alpha_i \neq 0$ à toutes les itérations.

En élaborant des algorithmes nous supposerons remplie (4.18). Une fois qu'un algorithme sera construit, nous vérifierons s'il en est bien ainsi et dans la négative nous imposerons des contraintes supplémentaires sur l'algorithme.

Ces remarques aidant, abordons directement les relations permettant de former des vecteurs A-conjugués.

Notons

$$r_i = x_{i+1} - x_i = \alpha_i p_i, \quad e_i = f'_{i+1} - f'_i = \alpha_i A p_i. \tag{4.19}$$

Prenons comme p_0 une direction de descente quelconque de la fonction (4.1), i.e. posons $p_0 = -H_0^* f'_0$, H_0 étant une matrice symétrique strictement définie positive. Etablissons les exigences que doit remplir le vecteur p_k, $1 \leqslant k \leqslant n-1$, pour être dans les conditions de A-conjugaison :

$$(p_k, Ap_j) = 0, \quad 0 \leqslant j \leqslant k-1. \tag{4.20}$$

Nous utiliserons le fait qu'en vertu des propriétés des directions conjuguées (voir (4.15)), si l'on choisit α_i dans (4.13) par la formule (4.16), on a, en plus des conditions (4.20), les égalités

$$(f'_k, p_j) = 0, \quad 0 \leqslant j \leqslant k-1. \tag{4.21}$$

Si l'on pose

$$p_k = -H_k^* f'_k, \tag{4.22}$$

H_k étant une matrice carrée $n \times n$, la condition (4.20) se récrit

$$(f'_k, H_k A p_j) = 0, \quad 0 \leqslant j \leqslant k-1.$$

La comparaison avec (4.21) montre que celle-ci implique (4.20) si la matrice H_k vérifie les relations

$$H_k A p_j = a p_j, \quad 0 \leqslant j \leqslant k-1,$$

avec a une constante quelconque.

Du moment qu'en vertu de la condition (4.18) et de la convexité stricte de (4.1) on a $0 < |\alpha_i| < \infty$ quel que soit $0 \leqslant i \leqslant n-1$, les égalités (4.20) et (4.21) s'écrivent, compte tenu de (4.19),

$$(r_k, e_j) = 0, \quad 0 \leqslant j \leqslant k-1, \tag{4.23}$$

$$(f'_k, r_j) = 0, \quad 0 \leqslant j \leqslant k-1, \tag{4.24}$$

et les conditions imposées à H_k prennent la forme

$$H_k e_j = a r_j, \quad 0 \leqslant j \leqslant k-1. \tag{4.25}$$

On a donc les conditions (4.20) si H_k définissant le vecteur p_k par la formule (4.22) vérifie les équations (4.25).

Si $k < n-1$, le nombre d'équations vectorielles (4.25) est inférieur à n et H_k se définit de façon multivoque. De plus, les systèmes d'équations donnant H_k diffèrent selon la valeur de la constante a. Cela atteste la diversité des algorithmes de construction de directions conjuguées parce que H_k différentes s'obtiennent par des méthodes distinctes.

Les équations (4.25) étant satisfaites pour tout $k = 1, 2, \ldots, n-1$, il est naturel de construire la matrice H_k à l'aide des relations de récurrence.

Mettons (4.25) sous la forme

$$(H_{k-1} + \Delta H_{k-1}) e_j = a r_j, \quad 0 \leqslant j \leqslant k-1. \tag{4.26}$$

Vu que la matrice H_{k-1} doit vérifier les équations

$$H_{k-1} e_j = a r_j, \quad 0 \leqslant j \leqslant k-2,$$

il résulte de (4.26) que ΔH_{k-1} est définie par les conditions

$$\Delta H_{k-1} e_j = 0, \quad 0 \leqslant j \leqslant k-2. \tag{4.27}$$

$$\Delta H_{k-1} e_{k-1} = a r_{k-1} - H_{k-1} e_{k-1}.$$

La dernière égalité a évidemment lieu si l'on pose

$$\Delta H_{k-1} = a\frac{r_{k-1}u_{k-1}^*}{(u_{k-1}, e_{k-1})} - \frac{H_{k-1}e_{k-1}v_{k-1}^*}{(v_{k-1}, e_{k-1})}, \tag{4.28}$$

avec u_{k-1}, v_{k-1} les vecteurs inconnus. On choisit ces vecteurs de façon à avoir la première condition (4.27), i.e. on a

$$(u_{k-1}, e_j) = 0, \quad (v_{k-1}, e_j) = 0, \quad 0 \leqslant j \leqslant k-2. \tag{4.29}$$

On conçoit qu'ils doivent également remplir les conditions suivantes :

$$(u_{k-1}, e_{k-1}) \neq 0, \quad (v_{k-1}, e_{k-1}) \neq 0. \tag{4.30}$$

En prenant en considération (4.23) on comprend que les conditions (4.29) sont remplies pour $u_{k-1} = v_{k-1} = r_{k-1}$. On a alors (4.30) puisque

$$(r_{k-1}, e_{k-1}) = (r_{k-1}, Ar_{k-1}) > 0 \tag{4.31}$$

conformément aux propriétés de la matrice A.

On peut également choisir u_{k-1}, v_{k-1} à partir des considérations suivantes. Si l'on a (4.20), alors

$$(Ap_{k-1}, p_j) = \left(\frac{1}{\alpha_{k-1}}e_{k-1}, \frac{1}{\alpha_j}r_j\right) = 0, \quad 0 \leqslant j \leqslant k-2.$$

Par suite, compte tenu de (4.25),

$$a(e_{k-1}, r_j) = (e_{k-1}, H_{k-1}e_j) = (H_{k-1}^*, e_{k-1}, e_j) = 0,$$
$$0 \leqslant j \leqslant k-2,$$

et, pour avoir (4.29), on peut poser

$$u_{k-1} = v_{k-1} = H_{k-1}^*e_{k-1}.$$

Si l'on choisit généralement les vecteurs en question sous la forme

$$u_{k-1} = t_{1,k}r_{k-1} + t_{2,k}H_{k-1}^*e_{k-1}, \tag{4.32}$$
$$v_{k-1} = t_{3,k}r_{k-1} + t_{4,k}H_{k-1}^*e_{k-1},$$

avec $t_{1,k}$, $t_{2,k}$, $t_{3,k}$, $t_{4,k}$ des nombres arbitraires (qui peuvent en principe varier avec k), on a évidemment les conditions (4.29). Pour avoir (4.30) on règle au besoin les quantités $t_{i,k}$, $i = 1, \ldots, 4$ (en particulier, ces conditions sont vérifiées, nous venons de le dire, pour $t_{1,k} = t_{3,k} = 1$, $t_{2,k} = t_{4,k} = 0$; voir (4.31)).

Le choix des vecteurs u_{k-1}, v_{k-1} sous forme (4.32) permet donc de générer la matrice ΔH_{k-1} par la formule (4.28) et de définir par là même des relations récurrentes pour construire une matrice H_k telle que le vecteur p_k qu'elle définit satisfasse aux conditions (4.20). Ceci étant, il correspond à tout couple de vecteurs u_{k-1}, v_{k-1} et à la constante a une matrice ΔH_{k-1} et donc H_k déterminées. En d'autres termes, selon les vecteurs u_k, v_k et la constante a choisis, on obtient divers algorithmes de construction de vecteurs conjugués, i.e. les différentes méthodes des directions conjuguées.

3. Propriétés générales des méthodes. Dégageons les propriétés générales dont jouissent les méthodes des directions conjuguées construites de manière décrite.

On élucide avant tout si, pour des méthodes étudiées, on a la condition (4.18) vu qu'en établissant les principes à la base de la formation des algorithmes cette condition est supposée remplie.

Une autre question intéressante à trancher concerne les directions p_j, $j = 0, 1, \ldots, n-1$, définies par H_j différentes. On veut précisément connaître si ces directions diffèrent, i.e. si les points $x_1, \ldots, x_{n-1}$ diffèrent selon l'algorithme (le point x_0 étant le même) ou s'ils coïncident.

Pour répondre à ces questions, utilisons la formule de récurrence pour la matrice H_j et les expressions (4.28), (4.32) et mettons le vecteur $-p_j = H_j^* f_j'$ sous la forme

$$H_j^* f_j' = (H_{j-1} + \Delta H_{j-1})^* f_j'.$$

Compte tenu des conditions (4.24),

$$\Delta H_{j-1}^* f_j' = -\frac{v_{j-1}(e_{j-1}, H_{j-1}^* f_j')}{(v_{j-1}, e_{j-1})} =$$

$$= -\frac{(t_{3,j} r_{j-1} + t_{4,j} H_{j-1}^* e_{j-1})(e_{j-1}, H_{j-1}^* f_j')}{(v_{j-1}, e_{j-1})}.$$

Si l'on se rappelle que

$$H_{j-1}^* e_{j-1} = H_{j-1}^* f_j' - H_{j-1}^* f_{j-1}' = H_{j-1}^* f_j' + p_{j-1},$$

le vecteur $-p_j$ se récrit

$$H_j^* f_j' = H_{j-1}^* f_j' \left[1 - \frac{t_{4,j}(e_{j-1}, H_{j-1}^* f_j')}{(v_{j-1}, e_{j-1})}\right] -$$

$$- \frac{t_{3,j} r_{j-1} e_{j-1}^* + t_{4,j} p_{j-1} e_{j-1}^*}{(v_{j-1}, e_{j-1})} H_{j-1}^* f_j' =$$

$$= \left[I\left(1 - \frac{t_{4,j}(e_{j-1}, H_j^* f_j')}{(v_{j-1}, e_{j-1})}\right) - \frac{r_{j-1} e_{j-1}^*}{(v_{j-1}, e_{j-1})}\left(t_{3,j} + \right.\right.$$

$$\left.\left. + \frac{t_{4,j}}{\alpha_{j-1}}\right)\right] H_{j-1}^* f_j'.$$

Ensuite,

$$(v_{j-1}, e_{j-1}) = (t_{3,j} r_{j-1} + t_{4,j} H_{j-1}^* e_{j-1}, e_{j-1}) = t_{3,j}(r_{j-1}, e_{j-1}) +$$

$$+ t_{4,j}(e_{j-1}, H_{j-1}^* f_j') + t_{4,j}(e_{j-1}, p_{j-1}), \tag{4.33}$$

ce qui fait que

$$\frac{(r_{j-1}, e_{j-1})}{(v_{j-1}, e_{j-1})}\left(t_{3,j} + \frac{t_{4,j}}{\alpha_{j-1}}\right) = 1 - \frac{t_{4,j}(e_{j-1}, H_{j-1}^* f_j')}{(v_{j-1}, e_{j-1})}$$

Usilisons cette expression pour la formule donnant $H_j^* f_j'$, il vient

$$H_j^* f_j' = \gamma_j \left(I - \frac{r_{j-1} e_{j-1}^*}{(r_{j-1}, e_{j-1})}\right) H_{j-1}^* f_j', \tag{4.34}$$

où

$$\gamma_j = 1 - \frac{t_{4,j}(e_{j-1}, H_{j-1}^* f_j')}{(v_{j-1}, e_{j-1})}.$$

Si le vecteur v_{j-1} vérifie la condition (4.30) et $\alpha_{j-1} t_{3,j} \neq -t_{4,j}$, alors, quel que soit $j = 1, 2, \ldots$, le facteur γ_j est différent de zéro parce que

$$\frac{t_{4,j}(e_{j-1}, H_{j-1}^* f_j')}{(v_{j-1}, e_{j-1})} \neq 1,$$

ce dont on se covainc facilement en comparant le numérateur avec l'expression (4.33). *Nous raisonnerons dans la suite en supposant les facteurs $t_{3,j}$ et $t_{4,j}$ choisis de façon qu'on ait $(v_{j-1}, e_{j-1}) \neq 0$ et $\gamma_j \neq 0$ pour $j \geqslant 1$.*

Multiplions scalairement par f_k' les deux membres de (4.34):

$$(f_k', H_j^* f_j') = \gamma_j \left[(f_k')^* - \frac{(f_k', r_{j-1}) e_{j-1}^*}{(r_{j-1}, e_{j-1})} \right] H_{j-1}^* f_j'. \tag{4.35}$$

Puisque $\gamma_j \neq 0$ et $(f_k', H_j^* f_j') = 0$, $(f_k', r_j) = 0$ pour $j \leqslant k - 1$ (en vertu de (4.21) et (4.24)), il découle de (4.35)

$$(f_k', H_{j-1}^* f_j') = 0, \quad 1 \leqslant j \leqslant k - 1. \tag{4.36}$$

Retranchons de (4.36) les égalités $(f_k', H_{j-1}^* f_{j-1}') = 0$, $1 \leqslant j \leqslant k - 1$, il vient également

$$(f_k', H_i^* e_i) = 0, \quad 0 \leqslant i \leqslant k - 2. \tag{4.37}$$

Utilisons les relations obtenues pour prouver qu'on a, pour $0 \leqslant i \leqslant j - 2$, les égalités $H_{i+1}^* f_j' = H_i^* f_j'$, i.e.

$$H_{j-1}^* f_j' = H_{j-2}^* f_j' = \ldots = H_0 f_j'. \tag{4.38}$$

Moyennant la formule récurrente pour la matrice H_i et les conditions (4.24) on récrit le vecteur $H_{i+1}^* f_j'$:

$$H_{i+1}^* f_j' = H_i^* f_j' - \frac{v_i (H_i e_i, f_j')}{(v_i, e_i)}, \quad 0 \leqslant i \leqslant j - 2. \tag{4.39}$$

Les égalités (4.38) se démontrent en faisant voir que

$$(f_j', H_i e_i) = 0, \quad 0 \leqslant i \leqslant j - 2. \tag{4.40}$$

Utilisons une fois de plus la formule récurrente pour H_i, il vient

$$(f_j', H_{i+1} e_{i+1}) = (f_j', H_0 e_{i+1}) - \sum_{s=0}^{i} \frac{(f_j', H_s e_s)(v_s, e_{i+1})}{(v_s, e_s)}, \tag{4.41}$$
$$0 \leqslant i \leqslant j - 2,$$

$$(f_j', H_{i+1}^* e_{i+1}) = (f_j', H_0 e_{i+1}) - \sum_{s=0}^{i} \frac{(v_s, f_j')(H_s e_s, e_{i+1})}{(v_s, e_s)}, \tag{4.42}$$
$$0 \leqslant i \leqslant j - 2.$$

En vertu des conditions (4.24) et (4.37),

$$(v_i, f'_j) = t_{3,i}(r_i, f'_j) + t_{4,i}(H_i^* e_i, f'_j) = 0, \quad 0 \leqslant i \leqslant j-2. \tag{4.43}$$

Vu ces égalités et les conditions (4.37), (4.42) entraîne

$$(f'_j, H_0 e_{i+1}) = 0, \quad 0 \leqslant i \leqslant j-3. \tag{4.44}$$

Analysons les relations (4.41). Pour $i = 0$, on a $(H_0 e_0, f'_j) = (H_0^* e_0, f'_j) = 0$ (en vertu de (4.37)) et $(f'_j, H_0 e_1) = 0$ (en vertu de (4.44)). Il se trouve donc que $(H_1 e_1, f'_j) = 0$. En poursuivant de la sorte on établit la validité des égalités (4.38).

Compte tenu de celles-ci, (4.34) se récrit

$$H_j^* f'_j = \gamma_j \left(I - \frac{r_{j-1} e_{j-1}^*}{(r_{j-1}, e_{j-1})} \right) H_0 f'_j, \tag{4.45}$$

formule qui permettra de répondre aux questions posées au début de ce numéro.

En multipliant (scalairement) par f'_j les deux membres de (4.45) et en tenant compte des conditions (4.24) on obtient

$$-(f'_j, p_j) = \gamma_j (f'_j, H_0 f'_j), \quad j \geqslant 0. \tag{4.46}$$

Si H_0 est strictement définie positive, alors $(f'_j, H_0 f'_j) > 0$. Par suite, si $\gamma_j \neq 0$, (4.46) entraîne $(f'_j, p_j) \neq 0$.

Ainsi, *l'égalité* (4.46) *entraîne que l'hypothèse sur la condition* (4.18), *hypothèse utilisée dans la construction des vecteurs conjugués, est juste à condition de prendre pour* H_0 *une matrice symétrique strictement définie positive.*

Pour voir si les vecteurs p_i et les points x_{i+1}, $i = 0, 1, \ldots, n-1$, diffèrent selon l'algorithme choisi, reprenons la formule (4.45).

Quelle que soit la méthode des directions conjuguées, le premier pas reste le même (pour une même H_0) puisque $x_1 = x_0 - \alpha_0 H_0^* f'_0$, et α_0 est choisi à partir de la condition $\min_\alpha f(x_0 - \alpha H_0^* f'_0)$. Par suite, le point x_1 et donc les vecteurs r_0, e_0, f'_1 sont les mêmes pour tout algorithme construit comme nous l'avons décrit. Il résulte alors de (4.45) que la direction

$$p_1 = -H_1^* f'_1 = -\gamma_1 \left[I - \frac{r_0 e_0^*}{(r_0, e_0)} \right] H_0 f'_1$$

ne dépend pas elle non plus du choix des vecteurs v_0, u_0 (vérifiant les exigences formulées), i.e. elle n'est pas fonction de la formation de la matrice H_1. Plus précisément, les vecteurs p_1 ainsi obtenus ne diffèrent que par le facteur scalaire γ_1. La quantité α_1 étant choisie par la condition $\min_\alpha f(x_1 + \alpha p_1)$, le point x_2 réalisant ce minimum est le même quel que soit le procédé de formation de H_1 (en vertu de la convexité stricte de $f(x)$). Par suite, c'est également la propriété des r_1, e_1, f'_2 : ces quantités restent les mêmes pour toute méthode des

directions conjuguées. En nous inspirant toujours de la représentation (4.45) du vecteur p_k nous verrons enfin que les points $x_0, x_1, \ldots, x_n$ sont indépendants du choix des vecteurs v_k, u_k, i.e. du procédé de construction de la matrice H_k. Ainsi, on obtient les mêmes approximations successives de la solution d'un problème de minimisation d'une fonction quadratique dans toutes les méthodes des directions conjuguées.

Signalons le fait suivant.

Nous avons remarqué que la première des deux matrices formant ΔH_j (4.28), $j = 0, 1, \ldots, k-1$, ne participe pas dans la construction de p_k. En effet, en définissant le vecteur $\Delta H_j^* f_k'$, on a, en vertu des conditions (4.24),

$$a\frac{u_j(r_j, f_k')}{(u_j, e_j)} = 0, \quad 0 \leqslant j \leqslant k-1,$$

i.e. les matrices

$$a\frac{r_j u_j^*}{(u_j, e_j)}, \quad 0 \leqslant j \leqslant k-1,$$

ne prennent pas part à la formation du vecteur $-p_k = H_k^* f_k' = (H_0 + \sum_{j=0}^{k-1} \Delta H_j)^* f_k'$. Elles n'en influent pas moins essentiellement sur les propriétés de la matrice H_k, en particulier sur celles de H_n. Nous en tiendrons compte au numéro suivant où nous étudierons des algorithmes concrets. Notons pour le moment que s'agissant de la recherche du minimum d'une fonction non quadratique, les propriétés des méthodes des directions conjuguées se ressentiront de la différence de propriétés des H_n.

4. Algorithmes concrets. Considérons plusieurs formules permettant de construire des directions conjuguées. Répétons que chacune d'elles détermine une méthode des directions conjuguées consistant à construire les approximations successives de la solution par les formules

$$x_{k+1} = x_k + \alpha_k p_k, \quad p_k = -H_k^* f_k', \quad k = 0, 1, \ldots, n-1, \tag{4.47}$$

α_k étant choisi à partir de la condition $\min_\alpha f(x_k + \alpha p_k)$, i.e. il est défini par les expressions (4.12) ou (4.16).

1) Posons dans (4.28) $a = 1$, $u_{k-1} = r_{k-1}$, $v_{k-1} = H_{k-1}^* e_{k-1}$ (i.e. dans (4.32) $t_{1,k} = t_1 = 1$, $t_{2,k} = t_2 = 0$, $t_{3,k} = t_3 = 0$, $t_{4,k} = t_4 = 1$). Alors

$$H_k = H_{k-1} + \frac{r_{k-1} r_{k-1}^*}{(r_{k-1}, e_{k-1})} - \frac{H_{k-1} e_{k-1} e_{k-1}^* H_{k-1}}{(H_{k-1}^* e_{k-1}, e_{k-1})}. \tag{4.48}$$

Etudions des propriétés de la matrice H_k ainsi obtenue.

La matrice H_k est symétrique. On l'établit facilement par récurrence. La matrice H_0 est symétrique. Il en est également des deux matrices formant ΔH_0 (la seconde l'est en vertu de la symétrie de H_0). Donc H_1 jouit elle aussi de cette propriété. On raisonne de même pour tout $k = 2, \ldots, n$.

La matrice H_k est strictement définie positive. Raisonnons par récurrence. La matrice H_0 est strictement définie positive. Soit H_k une matrice strictement définie positive. On a alors pour tout $x \in E^n$

$$(H_{k+1}x, x) = (H_k x, x) + \frac{(r_k, x)^2}{(r_k, e_k)^2} - \frac{(H_k e_k, x)^2}{(H_k e_k, e_k)} =$$
$$= \frac{(H_k x, x)(H_k e_k, e_k) - (H_k e_k, x)^2}{(H_k e_k, e_k)} + \frac{(r_k, x)^2}{(r_k, e_k)^2}.$$

Avec l'hypothèse sur H_k il existe une racine carrée $H_k^{1/2}$ (D. Faddeev et V. Faddeeva [57]). La matrice H_k étant symétrique, on a donc

$$(H_k x, x) = (H_k^{1/2} H_k^{1/2} x, x) = (H_k^{1/2} x, H_k^{1/2} x) = (y, y).$$

De même

$$(H_k e_k, e_k) = (H_k^{1/2} e_k, H_k^{1/2} e_k) = (z, z),$$
$$(H_k e_k, x) = (H_k^{1/2} e_k, H_k^{1/2} x) = (z, y).$$

Compte tenu de ces relations et de l'inégalité de Cauchy-Bouniakovski, on établit l'inégalité

$$(H_k x, x)(H_k e_k, e_k) - (H_k e_k, x)^2 = (y, y)(z, z) - (z, y)^2 \geqslant 0$$

l'égalité n'ayant lieu que si $z = y$, i.e., en vertu de la régularité de H_k, seulement pour $x = e_k$. D'autre part, $(r_k, x) = (r_k, e_k) = (r_k, Ar_k) > 0$. Ainsi, pour tout x différent de zéro,

$$(H_{k+1}x, x) = \frac{(y, y)(z, z) - (z, y)^2}{(H_k e_k, e_k)} + \frac{(r_k, x)^2}{(r_k, e_k)} > 0,$$

ce qui démontre la justesse de nos raisonnements.

La matrice $H_n = A^{-1}$. En effet, H_k satisfait aux équations (4.25) pour $a = 1$, i.e. $H_n e_j = r_j$, $j = 0, 1, \ldots, n-1$, ou, compte tenu de (4.19),

$$H_n A r_j = r_j, \quad j = 0, 1, \ldots, n-1.$$

Les vecteurs $r_0, \ldots, r_{n-1}$ sont donc vecteurs propres de la matrice $H_n A$ avec valeurs propres égales à l'unité. Vu l'indépendance linéaire des vecteurs conjugués r_i, $i = 0, 1, \ldots, n-1$, il en découle $H_n A = I$, i.e.

$$H_n = A^{-1}.$$

Mais (voir (4.8))

$$A^{-1} = \sum_{i=0}^{n-1} \frac{r_i r_i^*}{(Ar_i, r_i)} = \sum_i \frac{r_i r_i^*}{(e_i, r_i)},$$

i.e. H_n est définie par les seules matrices

$$\frac{r_i u_i^*}{(r_i, e_i)} = \frac{r_i r_i^*}{(r_i, e_i)}$$

(fait mentionné à la fin du numéro précédent).

2) Une autre méthode de construction de H_k consiste à poser $a = 1$ dans (4.28) et à choisir $u_{k-1} = v_{k-1} = r_{k-1}$. Alors

$$H_k = H_{k-1} + (r_{k-1} - H_{k-1}e_{k-1}) \frac{r_{k-1}^*}{(r_{k-1}, e_{k-1})}. \tag{4.49}$$

Cette fois, la matrice H_k n'est pas symétrique. Puisque $a = 1$, on a $H_n = A^{-1}$ (la démonstration coïncide avec celle pour la méthode (4.48)).

Moyennant (4.49) on construit une formule quelque peu différente pour définir H_k. Mettons (4.49) sous la forme

$$H_k = H_0 + \sum_{i=0}^{k-1} (r_i - H_i e_i) \frac{r_i^*}{(r_i, e_i)}. \tag{4.50}$$

En vertu des conditions de conjugaison (4.20) (compte tenu des formules (4.19)) on a $(e_k, r_j) = 0$, $0 \leqslant j \leqslant k - 1$. Par conséquent, (4.50) entraîne

$$H_k e_k = H_0 e_k, \quad k = 0, 1, \ldots, n - 1. \tag{4.51}$$

Etant donné cette égalité, la formule (4.49) se récrit

$$H_k = H_{k-1} + (r_{k-1} - H_0 e_{k-1}) \frac{r_{k-1}^*}{(r_{k-1}, e_{k-1})}. \tag{4.52}$$

Si $H_0 = I$, (4.52) s'avère plus simple que (4.49).

3) Prenons $a = 0$, $v_{k-1} = r_{k-1}$. Ceci étant,

$$H_k = H_{k-1} - \frac{H_{k-1} e_{k-1} r_{k-1}^*}{(r_{k-1}, e_{k-1})}. \tag{4.53}$$

(4.25) entraîne, pour $a = 0$, $H_n e_j = 0$, $j = 0, 1, \ldots, n - 1$. Etant donné l'indépendance linéaire des vecteurs $e_0, \ldots, e_{n-1}$, ces égalités signifient que $H_n = 0$ (l'indépendance linéaire des $e_i = Ar_i$, $i = 0, 1, \ldots, n - 1$, découle de celle des vecteurs conjugués r_i et des propriétés de la matrice A).

Puisque la condition (4.51) a également lieu pour (4.53), la dernière formule peut se mettre sous la forme

$$H_k = H_{k-1} - \frac{H_0 e_{k-1} r_{k-1}^*}{(r_{k-1}, e_{k-1})}. \tag{4.54}$$

On construit d'autres méthodes des directions conjuguées en combinant de diverses façons la constante a et les vecteurs u_k, v_k d'après les formules (4.32), mais nous nous en abstiendrons (en parlant

d'une méthode concrète, disons (4.48), nous entendrons dans la suite le procédé (4.47) où l'on construit la matrice H_k par la formule (4.48)).

Faisons une remarque. Strictement parlant, il faut, dans chacune des méthodes susmentionnées, vérifier les conditions (4.30) pour les vecteurs u_k, v_k. On s'assure facilement que ces conditions ont lieu dans tous les cas. Ainsi, on en a déjà parlé au n° 2 pour $u_k = v_k = r_k$. Dans la méthode (4.48) $v_k = H_k^* e_k$, mais la définie positivité de H_k entraîne $(v_k, e_k) = (H_k e_k, e_k) > 0$, i.e. les conditions (4.30) sont vérifiées. En conformité avec les résultats du n° 3 les méthodes considérées remplissent par conséquent la condition (4.18), i.e. on garantit leur non-dégénérescence.

Construisons les formules pour le calcul direct des vecteurs p_k définis par diverses matrices H_k. On y arrive sans peine à condition d'utiliser (4.45).

Puisque $r_{k-1} = \alpha_{k-1} p_{k-1}$, (4.45) entraîne

$$p_k = -\gamma_k (H_0 f'_k - \beta_k p_{k-1}), \tag{4.55}$$

où

$$\beta_k = \frac{(H_0 f'_k, e_{k-1})}{(p_{k-1}, e_{k-1})}. \tag{4.56}$$

Si $v_k = H_k^* e_k$, alors

$$\gamma_k = 1 - \frac{(e_{k-1}, H_{k-1}^*, f'_k)}{(H_{k-1} e_{k-1}, e_{k-1})}.$$

En vertu des égalités (4.38), $(e_{k-1}, H_{k-1}^* f'_k) = (e_{k-1}, H_0 f'_k)$. Ensuite, on a, comme le veulent (4.21) et (4.38),

$$(H_k e_k, e_k) = (H_k f'_{k+1}, f'_{k+1}) + (H_k f'_k, f'_k) = (H_0 f'_{k+1}, f'_{k+1}) - \\ - (p_k, f'_k).$$

Les égalités obtenues montrent que

$$\gamma_k = 1 - \frac{(H_0 f'_k, e_{k-1})}{(H_0 f'_k, f'_k) - (p_{k-1}, f'_{k-1})}. \tag{4.57}$$

Notons qu'en vertu de (4.21) et (4.24) il résulte de (4.45) (sous la condition $\gamma_j \neq 0$)

$$(f'_k, H_0 f'_j) = 0, \quad 0 \leqslant j \leqslant k-1. \tag{4.58}$$

Compte tenu de ce fait, on a

$$(H_0 f'_k, e_{k-1}) = (H_0 f'_k, f'_k). \tag{4.59}$$

Portons les égalités (4.59) dans (4.57), il vient

$$\gamma_k = -\frac{(p_{k-1}, f'_{k-1})}{(H_0 f'_k, f'_k) - (p_{k-1}, f'_{k-1})}. \tag{4.60}$$

Notons de plus que

$$(f'_k, p_k) = (f'_k, p_k) - (f'_{k+1}, p_k) = -(e_k, p_k). \tag{4.61}$$

En comparant les formules (4.56) et (4.60) et en tenant compte de (4.59) et (4.61) on établit aisément que $\gamma_k = 1/(1 + \beta_k)$. D'où $\gamma_k\beta_k = 1 - \gamma_k$. Par conséquent, la formule (4.55), donnant le vecteur p_k dans le cas où on construit la matrice H_k à l'aide de $v_{k-1} = H^*_{k-1}e_{k-1}$, peut se récrire comme

$$p_k = -\gamma_k H_0 f'_k + (1 - \gamma_k)\, p_{k-1}, \tag{4.62}$$

le coefficient γ_k étant défini soit par (4.57) soit par (4.60). Le vecteur p_k admet également l'écriture suivante:

$$p_k = -H_0 f'_k + \tilde{\beta}_k\, (H_0 f'_k + p_{k-1}), \tag{4.63}$$

où

$$\tilde{\beta}_k = 1 - \gamma_k = \frac{(H_0 f'_k, f'_k)}{(H_0 f'_k, f'_k) - (p_{k-1}, f'_{k-1})}. \tag{4.64}$$

D'autres expressions de $\tilde{\beta}_k$ sont également possibles si l'on recourt à (4.59), (4.61), (4.46); notons que la dernière formule peut s'écrire

$$(f'_k, p_k) = (\tilde{\beta}_k - 1)\, (H_0 f'_k, f'_k). \tag{4.65}$$

En employant diverses formules pour calculer les coefficients γ_k, $\tilde{\beta}_k$ dans (4.62), (4.63), nous aboutirons à autant de méthodes des directions conjuguées. Soulignons que dans la recherche du minimum de fonctions non quadratiques, diverses formules donnant p_k déterminent des vecteurs différents (pour ce qui est de leur grandeur et direction).

La construction la plus simple de p_k moyennant l'expression (4.45) est celle quand on forme la matrice H_k à l'aide du vecteur $v_{k-1} = r_{k-1}$. Alors $t_{4,k} = 0$, donc $\gamma_k = 1$, et (4.55) fournit

$$p_k = -H_0 f'_k + \beta_k p_{k-1}, \tag{4.66}$$

avec β_k trouvé d'après la formule (4.56). Si l'on utilise les égalités (4.59), (4.61) et (4.46), cette dernière s'écrivant dans le cas considéré

$$(p_k, f'_k) = -(H_0 f'_k, f'_k), \tag{4.67}$$

les coefficients β_k se définissent par exemple par l'une des formules suivantes:

$$\beta_k = -\frac{(H_0 f'_k, e_{k-1})}{(p_{k-1}, f'_{k-1})} = -\frac{(H_0 f'_k, f'_k)}{(p_{k-1}, f'_{k-1})} = \frac{(H_0 f'_k, f'_k)}{(H_0 f'_{k-1}, f'_{k-1})}. \tag{4.68}$$

Les expressions (4.62), (4.63), (4.66) donnant le vecteur p_k se mettent à leur tour sous la forme $p_k = -H^*_k f'_k$, la matrice H_k dépendant de la forme des coefficients γ_k, $\tilde{\beta}_k$, β_k. Ainsi, si le coefficient

$\tilde{\beta}_k$ de (4.63) se calcule par la formule (4.64), la matrice correspondante

$$H_k = H_0 - \frac{H_0 f'_k (H_0 f'_k + p_{k-1})^*}{(H_0 f'_k, f'_k) - (p_{k-1}, f'_{k-1})}. \tag{4.69}$$

Ici $H_n = H_0$ (du moment que $f'_n = 0$). Si dans (4.66) β_k s'obtient par la première formule (4.68), le vecteur p_k est défini par la matrice

$$H_k = H_0 + \frac{H_0 e_{k-1} \cdot p^*_{k-1}}{(p_{k-1}, f'_{k-1})}. \tag{4.70}$$

Si β_k de (4.66) est calculé moyennant la seconde formule (4.68), alors

$$H_k = H_0 + \frac{H_0 f'_k p^*_{k-1}}{(p_{k-1}, f'_{k-1})}. \tag{4.71}$$

Notons que dans (4.71) $H_n = H_0$ (vu que $f'_n = 0$), et dans (4.70) $H_n \neq H_0$.

S'il en a le désir, le lecteur peut établir facilement d'autres formules donnant H_k.

La plus simple formule de calcul de vecteurs A-conjugués s'obtient en choisissant dans (4.66) $H_0 = I$. Alors

$$p_k = -f'_k + \beta_k p_{k-1}, \tag{4.72}$$

avec β_k défini par exemple par l'une des formules suivantes:

$$\beta_k = -\frac{(f'_k, e_{k-1})}{(p_{k-1}, f'_{k-1})} = -\frac{(f'_k, f'_k)}{(p_{k-1}, f'_{k-1})} = \frac{(f'_k, f'_k)}{(f'_{k-1}, f'_{k-1})}. \tag{4.73}$$

La méthode (4.47) où la construction de vecteurs conjugués s'effectue selon (4.72), (4.73) est connue sous le nom de méthode du gradient conjugué (nom qui découle des conditions (4.58)).

5. Minimisation d'une fonction quadratique convexe. Les méthodes des directions conjuguées ne se sont appliquées jusqu'à présent qu'à des fonctions quadratiques strictement convexes, i.e. la matrice A a été supposée strictement définie positive.

Soit

$$f(x) = \frac{1}{2}(Ax, x) + (b, x) + c$$

une fonction convexe, i.e. A est définie positive: $(Ax, x) \geqslant 0$ pour tout $x \neq 0$. Supposons que $f(x)$ admet bien un minimum.

Voyons comment jouent dans ce cas les méthodes des directions conjuguées. Stipulons d'abord certaines propriétés de la fonction $f(x)$.

1) *Si* $(Ap, p) = 0$, *on a nécessairement*

$$Ap = 0. \tag{4.74}$$

En effet, sous l'hypothèse faite p est un point de minimum de la fonction convexe $\varphi(x) = \frac{1}{2}(Ax, x)$. Or, le point de minimum doit vérifier une condition nécessaire d'extrémum

$$\varphi'(p) = Ap = 0.$$

2) *Si p est un point de minimum de la fonction convexe $\varphi(x) = \frac{1}{2}(Ax, x)$, on a nécessairement*

$$(b, p) = 0. \tag{4.75}$$

En effet, si $(Ap, p) = 0$ et $(b, p) > 0$, alors $f(\alpha p) = \alpha(b, p) + c \to -\infty$ pour $\alpha \to -\infty$, i.e. $f(x)$ n'admet pas de minimum: contradiction avec l'hypothèse ititiale.

Le cas $(b, p) < 0$ se traite de façon analogue.

3) *La fonction $f(x)$ possède plus d'un point de minimum.*

En effet, tout point de minimum de la fonction quadratique convexe $f(x)$ doit être solution du système linéaire $Ax + b = 0$ et vice versa parce que $f'(x) = Ax + b = 0$ est une condition nécessaire et (vu l'existence d'un minimum de $f(x)$) suffisante d'extrémum de la fonction convexe $f(x)$ (conséquence I.3.2). Le rang de la matrice A étant inférieur au nombre d'inconnues (la condition $(Ax, x) \geqslant 0$ signifie la singularité de A, voir (4.74)), le système $Ax + b = 0$ a plus d'une solution.

4) *Si $(Ap, p) = 0$ et $z \in E^n$ un point quelconque, on a nécessairement*

$$(f'(z), p) = 0. \tag{4.76}$$

En effet, en vertu des conditions (4.74) et (4.75)

$$(f'(z), p) = (Az + b, p) = (Ap, z) + (b, p) = 0.$$

L'égalité (4.76) s'interprète comme suit. L'ensemble de solutions du problème de minimisation de $\varphi(x)$ forme un hyperplan de dimension $n-q$, où q est le rang de A. (Cet hyperplan appartient à une surface de niveau de $f(x)$ parce que si p est un point de minimum arbitraire de $\varphi(x) = \frac{1}{2}(Ax, x)$, alors, compte tenu de (4.75),

$$f(p) = \frac{1}{2}(Ap, p) + (b, p) + c = c.)$$

L'égalité (4.76) signifie donc que le gradient de $f(x)$ en tout point est dans un sous-espace de dimension q orthogonal au plan $Ap = 0$. Il en résulte que le nombre de vecteurs $f'(x)$ linéairement indépendants est $q < n$ (pour $f(x)$ quadratique convexe).

Forts de la connaissance des propriétés de $f(x)$ étudions les méthodes des directions conjuguées pour le problème considéré.

Supposons pour simplifier que $H_0 = I$ et abordons la méthode (4.72). Désignons par E^q le sous-espace d'éléments $f'(x)$. On s'assure

facilement que le vecteur p_k défini par la formule (4.72) appartient à E^q. En effet, $p_0 = -f'_0 \in E^q$ et par suite, quel que soit k, p_k est combinaison linéaire de vecteurs $\in E^q$. En utilisant la méthode (4.72) la minimisation de la fonction (processus (4.47)) a donc lieu en fait dans ce sous-espace. Or, celui-ci vérifie la condition $(Ax, x) > 0$ pour tout $x \neq 0$. Etant donné la dimension finie de E^q, cela signifie que, quel que soit $x \in E^q$, on a

$$m_1 \| x \|^2 \leqslant (Ax, x) \leqslant M_1 \| x \|^2, \quad m_1 > 0, \quad M_1 \leqslant M.$$

D'où la convexité stricte de $f(x)$ sur le sous-espace considéré et, partant, la validité de toutes les propriétés des méthodes des directions conjuguées que nous avons examinées aux numéros précédents. Ont toujours lieu, en particulier, les égalités (4.58) montrant l'indépendance linéaire des vecteurs f'_i, $i = 0, 1, \ldots, k$. Mais il existe dans E^q q vecteurs linéairement indépendants au plus. Par suite, le processus de construction de vecteurs conjugués s'arrête pour un certain $k \leqslant q - 1$. La méthode étant non dégénérée (propriété (4.18)), cela n'a lieu que lorsque $f'_k = 0$.

Nos raisonnements nous autorisent à dire qu'en minimisant par la méthode (4.72) pour un certain $k \leqslant q - 1$ on aura nécessairement $f'_k = 0$. Puisque les directions p_k définies par différentes méthodes coïncident (à un facteur scalaire près), tout ce que nous avons dit reste vrai pour les autres algorithmes étudiés dans le paragraphe 4. A force de compliquer un peu les développements, on montre qu'il en est également dans le cas où H_0 est strictement définie positive quelconque.

Ainsi, *les méthodes des directions conjuguées permettent la recherche du point de minimum d'une fonction quadratique convexe, et cela en un nombre de pas inférieur à n.*

Supposons cette fois que $f(x)$ convexe n'atteint pas son minimum ; il découle de la démonstration de l'égalité (4.75) que cela a lieu si un vecteur p quelconque réalisant le minimum de $\varphi(x) = (Ax, x)$ satisfait à la condition $(b, p) \neq 0$.

Voyons les résultats de la construction de directions conjuguées dans ce cas.

Donnons au préalable une propriété des vecteurs A-conjugués dont ceux-ci jouissent si $f(x)$ est convexe (sans l'être strictement).

Si A est définie positive, alors l'un au moins des vecteurs conjugués p_k, $0 \leqslant k \leqslant n - 1$, vérifie la condition

$$(Ap_k, p_k) = 0. \tag{4.77}$$

En effet, s'il y avait

$$(Ap_i, p_i) > 0 \tag{4.78}$$

pour tous les $i = 0, 1, \ldots, n - 1$, le système $p_0, p_1, \ldots, p_{n-1}$ serait linéairement indépendant. Supposons remplies les conditions (4.78)

et soit

$$\sum_{i=0}^{n-1} \delta_i p_i = 0,$$

avec $\delta_0 \neq 0$ par exemple. En multipliant (scalairement) par Ap_0 les deux membres on a à gauche $\delta_0 (Ap_0, p_0) > 0$: contradiction. Par conséquent, sous l'hypothèse de (4.78) vérifiée, les vecteurs $\{p_i\}_0^{n-1}$ formeraient une base dans E^n et tout vecteur z se mettrait sous la forme

$$z = \sum_{i=0}^{n-1} a_i p_i,$$

où l'un au moins des coefficients $a_i \neq 0$. On aurait alors

$$(Az, z) = \left(A \sum_i a_i p_i, \ \sum_i a_i p_i\right) = \sum_i a_i^2 (Ap_i, p_i) > 0.$$

Ce qui contredit la condition $(Ax, x) \geqslant 0$.

Nous n'avons donc été dans le vrai en supposant les conditions (4.78) satisfaites.

En vertu de la propriété examinée des vecteurs A-conjugués, dans toute méthode des directions conjuguées on a pour un certain $k \geqslant 0$ l'égalité (4.77). Comme la fonction considérée ne possède pas les propriétés (4.75), (4.76), le paramètre α_k de (4.47) calculé par la formule (4.16) devient infini si (4.77) est satisfait, i.e. il est impossible de poursuivre la construction de directions conjuguées.

6. Discussion des résultats. Ainsi, nous avons examiné un schéma général de construction des méthodes des directions conjuguées qui nous a fourni de nombreux algorithmes concrets. Toute méthode de la classe considérée permet de minimiser une fonction quadratique convexe en un nombre de pas de (4.47) égal au plus à n. Nous avons également constaté l'identité des approximations successives x_0, $x_1, \ldots, x_k$ de la solution dans divers algorithmes.

En appréciant ceux-ci du point de vue du volume de calcul par itération, c'est à (4.62), (4.63), (4.66) qu'on doit donner la préférence. Leur mise en œuvre est particulièrement simple si l'on prend pour matrice initiale H_0 la matrice unité I, et l'on voit que ce choix est justement le plus rationnel dans la plupart des problèmes.

Dans ce cas, le coût par itération des méthodes (4.63), (4.66) n'est que fort peu différent de celui de la méthode du gradient, mais, le processus (4.47) aboutissant au bout d'un nombre fini de pas, elles la dépassent de beaucoup sur le plan efficacité.

L'avantage des procédés (4.62), (4.63) et (4.66) est également un faible accroissement d'occupation de mémoire par rapport à la méthode de la plus grande pente.

Sous les aspects considérés, les méthodes des directions conjuguées dans lesquelles la direction de déplacement est définie après la construction de la matrice (méthodes (4.48), (4.49), (4.52)-(4.54)) le cèdent un peu à (4.63), (4.66) tout en étant notablement supérieures aux méthodes du gradient. Toutes les méthodes de la classe considérée sont meilleures que celle de Newton du fait de se passer du calcul des dérivées secondes.

Une question peut se poser : vaut-il la peine de s'occuper de méthodes avec construction préalable de la matrice qui le cèdent aux techniques (4.63), (4.72) quant au coût par itération et à l'encombrement de mémoire? Le lecteur est cependant prié de ne pas oublier que notre évaluation des méthodes est purement théorique et que nous ne tenons pas compte de ce facteur important qu'est la sensibilité des algorithmes aux erreurs de calcul. Or, si l'on en tenait compte, il se pourrait que les coûts relatifs de divers algorithmes changent notablement. Il faut dire de plus que dans les méthodes (4.48), (4.49), (4.52) on trouve le minimum et on obtient simultanément la matrice inverse A^{-1}, ce qui peut s'avérer utile dans certains cas.

C'est surtout en minimisant des fonctions non quadratiques qu'on se ressent de la différence de propriétés des algorithmes (voir paragraphe suivant).

Les méthodes des directions conjuguées rendent un autre service : elles permettent d'établir si la matrice est définie à signe fixe. Ainsi, selon les résultats du n° 5, si A est définie positive et $f(x)$ n'atteint pas son minimum, on a $\alpha_k = \infty$ à un pas. Et si A n'est pas définie positive, à un pas du processus (4.47) $\alpha_k < 0$. C'est donc la grandeur du paramètre α_k qui autorise à parler de la propriété en question de la matrice A.

L'efficacité des méthodes des directions conjuguées est la raison de leur utilisation de plus en plus grande dans la recherche du minimum de fonctions quadratiques et la résolution des systèmes d'équations linéaires.

§ 5. Méthodes des directions conjuguées. Minimisation de fonctions quelconques

1. Applicabilité des méthodes. Supposons que nous nous proposons de nous servir du processus

$$x_{k+1} = x_k + \alpha_k p_k, \quad p_k = -H_k^* f_k', \quad k = 0, 1, \ldots \tag{5.1}$$

avec p_k (ou H_k) défini par l'une des méthodes du paragraphe précédent, pour minimiser $f(x)$, une fonction convexe (non quadratique) quelconque. Les éléments de la matrice $f''(x)$ varient alors selon les points de la suite (5.1) et en construisant les vecteurs $p_0, \ldots, p_k$ par n'importe quel procédé du n° 4 § 4 on constate qu'ils ne vérifient plus les conditions (4.20), i.e. ils ne sont pas conjugués. Mais si le

point initial x_0 est choisi dans un voisinage suffisamment petit du point de minimum x_* d'une fonction convexe régulière $f(x)$, en tout point de ce voisinage la matrice $f''(x)$ est proche de $f''(x_*)$, i.e. la fonction de départ $f(x)$ est bien approchée par la fonction quadratique

$$\varphi(x) = \frac{1}{2}(f''(x_*)(x - x_*),\ x - x_*) + f(x_*).$$

On s'attend par suite à ce que les propriétés des vecteurs $p_0, \ldots, p_k$ définis par les méthodes du § 4 soient proches de celles des vecteurs conjugués ($f''(x_*)$-conjugués) et les propriétés du processus (5.1) où le paramètre α_k est choisi à partir de la condition de minimum de $f(x)$ dans la direction p_k s'apparentent à celles des méthodes des directions conjuguées que nous avons étudiées. En d'autres termes, on est autorisé à supposer les méthodes du paragraphe précédent suffisamment puissantes lorsqu'il s'agira de minimiser des fonctions non quadratiques. Elles ne possèdent certes plus la propriété de fournir la solution au bout d'un nombre fini de pas parce que les conditions

$$(f''(x_*)\, p_k,\ p_i) = 0, \quad i \neq k,$$

ne seront strictement vérifiées que pour un certain choix du point x_0.

Comme par le passé nous appellerons méthodes des directions conjuguées des processus itératifs du type (5.1) dans lesquels le vecteur p_k est construit par les algorithmes du § 4 et la valeur du paramètre α_k est choisie à partir de la condition

$$f(x_k + \alpha_k p_k) = \min_{\alpha} f(x_k + \alpha p_k).$$

Cette dernière condition peut également s'écrire

$$(f'_{k+1},\ p_k) = (f'(x_k + \alpha_k p_k),\ p_k) = 0. \tag{5.2}$$

Le présent paragraphe se propose de justifier la convergence des méthodes des directions conjuguées lors de la minimisation de fonctions non quadratiques et d'évaluer sa vitesse.

2. Théorème sur la convergence des méthodes. Nous supposerons dans la suite que $f(x)$ est une fonction deux fois continûment dérivable fortement convexe, i.e. on a

$$m \| y \|^2 \leqslant (f''(x)\, y,\ y) \leqslant M \| y \|^2, \quad m > 0, \tag{5.3}$$

quels que soient $x,\ y \in E^n$, et on choisit comme H_0 une matrice symétrique strictement définie positive

$$m_0 \| y \|^2 \leqslant (H_0 y,\ y) \leqslant M_0 \| y \|^2, \quad m_0 > 0, \tag{5.4}$$

pour tous les $y \in E^n$.

Les processus du type (5.1) ont lieu avec ou sans reconstitution de H_k au bout d'un nombre fini de pas. En parlant de processus avec reconstitution, disons en n pas, nous insinuons que, quel que soit $\xi = 0, 1, \ldots$, on reconstitue $H_{\xi n}$, i.e. $H_{\xi n} = H_0$.

Il faut dire de suite que s'agissant de processus avec reconstitution de H_k après un nombre fini de pas, toute méthode des directions conjuguées vérifie la condition

$$\lim_{k\to\infty} \| f'(x_k) \| = 0 \tag{5.5}$$

puisque chaque pas immédiatement consécutif à la reconstitution est un pas du gradient qui vérifie, en vertu de (5.3), les conditions de convergence des méthodes du gradient (théorème 1.6), et aux pas suivants on descend jusqu'au minimum de la fonction dans la direction de déplacement. La condition (5.5) pour une fonction strictement convexe signifie que *toute méthode du § 4 avec reconstitution de H_k en un nombre fini de pas converge vers la solution x_**. Il importe donc, lorsqu'il s'agit de se prononcer sur l'efficacité de processus avec reconstitution, d'évaluer leur vitesse de convergence.

Notons que dans le cas de tels processus la condition (5.5) est remplie non seulement pour des fonctions strictement convexes, mais aussi pour toute fonction telle que cette condition soit garantie dans les méthodes du gradient (voir théorème 1.4).

S'il n'y a pas reconstitution de H_k, on doit légitimer la convergence et, comme pour des processus avec reconstitution, évaluer sa vitesse.

Le théorème ci-dessus représente le résultat principal du présent paragraphe.

THÉORÈME 5.1. *Etant donné une fonction $f(x)$ vérifiant* (5.3) *qu'on minimise par le processus* (5.1) *avec H_k construite moyennant l'une des méthodes du* § 4 ((4.48)-(4.49), (4.52)-(4.54), (4.69)-(4.71)), *la reconstitution de H_k ayant lieu en n pas, si la valeur de α_k est choisie à partir de la condition de minimum dans la direction p_k, la convergence de la suite $\{x_{\xi n}\}$ vers la solution est superlinéaire quel que soit le point initial x_0.*

Donnons les grandes lignes de la démonstration. Supposons fausse l'affirmation du théorème, i.e. les processus itératifs décrits remplissent pour tout k la condition

$$\| x_{k+1} - x_* \| \geqslant \lambda \| x_k - x_* \|, \tag{5.6}$$

où $\lambda > 0$ est une constante. En utilisant les inégalités (1.12) et

$$\| f'(x) \| = \| f'(x) - f'(x_*) \| \leqslant M \| x - x_* \|, \tag{5.7}$$

valables pour une fonction vérifiant (5.3), on se convainc de l'équivalence de (5.6) et de

$$\| f'_{k+1} \| \geqslant \delta \| f'_k \|, \tag{5.8}$$

avec $\delta > 0$ une constante. Sous l'hypothèse d'avoir (5.8) nous établissons que pour le processus (5.1) on a, quel que soit l'algorithme de construction de H_k, les estimations

$$C \| f'_k \| \leqslant \| r_k \| \leqslant N \| f'_k \|, \tag{5.9}$$

avec C, N des constantes indépendantes de k, $C > 0$, et

$$(e_{\xi n+i}, r_{\xi n+j}) = o(\| e_{\xi n+i} \| \| r_{\xi n+j} \|), \quad i \neq j, \quad 0 \leqslant i, \ j \leqslant n-1. \tag{5.10}$$

Nous montrons ensuite qu'en présence de ces estimations la convergence de la suite (5.1) vers la solution est superlinéaire. Mais cela contredit l'hypothèse de départ (5.6) (ou (5.8)), i.e. le processus (5.1) ne peut pas satisfaire à la condition (5.6). L'affirmation du théorème s'en déduit facilement.

Le schéma de démonstration est donc le même pour toutes les méthodes étudiées. N'est différente que la manière dont on établit les estimations (5.9) et (5.10). Au numéro suivant nous prouverons celles-ci pour divers algorithmes en nous bornant pour le moment à exposer la démonstration commune à toutes les méthodes.

Faisons d'abord une remarque à propos des notations. Lorsque nous travaillerons avec les vecteurs et les paramètres

$$r_{\xi n+i}, \ f'_{\xi n+i}, \ e_{\xi n+i}, \ \alpha_{\xi n+i}, \ \beta_{\xi n+i},$$
$$i = 0, 1, \ldots, n-1,$$

nous omettrons souvent, pour alléger l'écriture, l'indice ξn, i.e. nous aurons affaire avec r_i, f'_i, e_i, α_i, β_i, et ainsi de suite. Soulignons que nous ne le faisons que pour simplifier l'écriture et qu'en fait les vecteurs correspondants restent indicés par $\xi n + i$.

Démontrons donc notre théorème en supposant vérifiées (5.9) et (5.10). Moyennant la formule de Lagrange généralisée nous obtenons

$$(e_i, \ r_j) = (f''_{ic} r_i, \ r_j) = (f''_i r_i, \ r_j) + ((f''_{ic} - f''_i) r_i, \ r_j). \tag{5.11}$$

L'indice ic désigne comme toujours un point intermédiaire du segment correspondant :

$$x_{ic} = x_i + \theta r_i, \quad 0 \leqslant \theta \leqslant 1.$$

Si $\| r_i \| \to 0$, vu la continuité uniforme des dérivées secondes de $f(x)$ sur l'ensemble $S = \{x : f(x) \leqslant f(x_0)\}$, on a $\| f''_{ic} - f''_i \| \to 0$ et (5.11) entraîne les estimations

$$(f''_i r_i, \ r_j) = o(\| r_i \| \| r_j \|) + o(\| e_i \| \| r_j \|),$$
$$i \neq j, \quad 0 \leqslant i, \quad j \leqslant n-1,$$

lorsque (5.10) a lieu. Dans les conditions (5.3) $\| e_i \| = \| f'_{i+1} - f'_i \| \leqslant \leqslant M \| r_i \|$, par suite $\| e_i \|$ et $\| r_i \|$ sont du même ordre infinitésimal. Compte tenu de ce fait,

$$(f''_i r_i, \ r_j) = o(\|r_i\| \ \|r_j\|), \quad i \neq j, \quad 0 \leqslant i, \ j \leqslant n-1. \tag{5.12}$$

Si l'on a (5.12), il existe des vecteurs

$$\bar{r}_i = r_i + \omega_i, \quad i = 0, 1, \ldots, n-1, \tag{5.13}$$

avec $\|\omega_i\| = o(\|r_i\|)$, tels que

$$(f''_{\xi n}\bar{r}_i, r_j) = 0, \ i \neq j, \ 0 \leqslant i, j \leqslant n-1. \tag{5.14}$$

Pour le montrer procédons comme suit. Normons les vecteurs r_i:

$$\tilde{r}_i = \frac{r_i}{(f''_{\xi n}r_i, r_i)^{1/2}}.$$

Alors $(f''_{\xi n}\tilde{r}_i, \tilde{r}_i) = 1$ et on a pour $\xi \to \infty$, en vertu de la convergence du processus (5.1) (avec reconstitution de H_k) et des conditions (5.3) et (5.12):

$$(f''_{\xi n}\tilde{r}_i, \tilde{r}_j) \leqslant \frac{1}{m\|r_i\|\|r_j\|}[(f''_i r_i, r_j) + ((f''_{\xi n} - f''_i)r_i, r_j)] \to 0,$$
$$i \neq j, \ 0 \leqslant i, \ j \leqslant n-1.$$

Si $\tilde{R}_{\xi n}$ est donc une matrice ayant pour colonnes les vecteurs $\tilde{r}_i$ et $F_{\xi n} = \tilde{R}^*_{\xi n} f''_{\xi n}\tilde{R}_{\xi n}$, alors, quand $\xi \to \infty$,

$$F_{\xi n} \to I.$$

Comme $F^{-1}_{\xi n}\tilde{R}^*_{\xi n} f''_{\xi n}\tilde{R}_{\xi n} = I$, on obtient

$$Q^*_{\xi n} f''_{\xi n}\tilde{R}_{\xi n} = I \tag{5.15}$$

en notant $F^{-1}_{\xi n}\tilde{R}^*_{\xi n} = Q^*_{\xi n}$. Mais, étant donné $F_{\xi n} \to I$, on a également $F^{-1}_{\xi n} \to I$ et donc $\tilde{R}^*_{\xi n} \to Q^*_{\xi n}$, i.e. les vecteurs-colonnes q_i de la matrice $Q_{\xi n}$ s'écrivent

$$q_i = \tilde{r}_i + \tilde{\omega}_i, \quad i = 0, 1, \ldots, n-1,$$

où $\|\tilde{\omega}_i\| \to 0$ pour ξ tendant vers l'infini. Mettons les égalités obtenues sous la forme

$$(f''_{\xi n}r_i, r_i)^{1/2} q_i = r_i + (f''_{\xi n}r_i, r_i)^{1/2}\tilde{\omega}_i.$$

En vertu de (5.15) les vecteurs r_i et $\bar{r}_i = (f''_{\xi n}r_i, r_i)^{1/2} q_i$, $i = 0, 1, \ldots$ $\ldots, n-1$, vérifient les conditions (5.14). D'autre part, $\bar{r}_i$ satisfont aux conditions (5.13) puisqu'on a, vu (5.3),

$$\frac{\|\omega_i\|}{\|r_i\|} = \frac{(f''_{\xi n}r_i, r_i)^{1/2}\|\tilde{\omega}_i\|}{\|r_i\|} \leqslant \frac{M^{1/2}\|r_i\|}{\|r_i\|}\|\tilde{\omega}_i\| \to 0.$$

Nous avons donc établi la validité de (5.14).

Pour ξ suffisamment grands, les vecteurs $\bar{r}_i$ sont linéairement indépendants. En effet, supposons l'existence de facteurs δ_i, $i = 0, 1, \ldots, n-1$, dont deux au moins sont non nuls et tels que

$\sum_{i=0}^{n-1} \delta_i \bar{r}_i = 0$. Si $\delta_0 \neq 0$, on a alors

$$\delta_0 (f''_{\xi n} r_0, \bar{r}_0) + \sum_{j=1}^{n} \delta_j (f''_{\xi n} r_0, \bar{r}_j) = 0.$$

Cette égalité n'a cependant pas lieu pour des ξ suffisamment importants. En effet, vu que $\| \omega_i \| = o(\| r_i \|)$ et $\| r_i \| \to 0$ quand $\xi \to \infty$, on a, compte tenu de (5.3), pour ξ suffisamment grands

$$(f''_{\xi n} r_0, \bar{r}_0) = (f''_{\xi n} r_0, r_0) + (f''_{\xi n} r_0, \omega_0) \neq 0,$$

alors que $(f''_{\xi n} r_0, \bar{r}_j) = 0$, $j = 1, \ldots, n-1$, en vertu de (5.14). Nous avons abouti à une contradiction, ce qui prouve l'indépendance linéaire des vecteurs $\bar{r}_i$, $i = 0, 1, \ldots, n-1$.

Soit $z_{\xi n}$ un point de minimum de la fonction quadratique

$$\varphi(x) = (f'_{\xi n}, x - x_{\xi n}) + \frac{1}{2} (f''_{\xi n} (x - x_{\xi n}), x - x_{\xi n}).$$

Ecrivons le vecteur $z_{\xi n} - x_{\xi n}$ comme

$$z_{\xi n} - x_{\xi n} = \sum_{i=0}^{n-1} a_i \bar{r}_i. \tag{5.16}$$

Comme $\varphi'(z_{\xi n}) = f'_{\xi n} + f''_{\xi n} (z_{\xi n} - x_{\xi n}) = 0$, on a, en utilisant (5.16),

$$\sum_{i=0}^{n-1} a_i f''_{\xi n} \bar{r}_i = -f'_{\xi n},$$

d'où, compte tenu de (5.14), la formule suivante pour calculer a_i:

$$a_i = -\frac{(f'_{\xi n}, r_i)}{(f''_{\xi n} \bar{r}_i, r_i)}, \quad i = 0, 1, \ldots, n-1.$$

Mettons le numérateur sous la forme

$$(f'_{\xi n}, r_i) = (f'_0 - f'_1 + f'_1 - \ldots - f'_{i+1} + f'_{i+1}, r_i) = -\sum_{j=0}^{i} (e_j, r_i)$$

(on prend en considération que $(f'_{i+1}, r_i) = 0$ en vertu de (5.2)). D'où, compte tenu des estimations (5.10),

$$(f'_{\xi n}, r_i) = -(e_i, r_i) + \sum_{j=0}^{i-1} o(\| r_i \| \, \| e_j \|). \tag{5.17}$$

Etant donné les conditions (5.8) et (5.9), tous les vecteurs $r_0, \ldots, r_{n-1}$ sont du même ordre infinitésimal (rappelons qu'il s'agit en fait des vecteurs $r_{\xi n+i}$). Vu que $\| e_i \| \leqslant M \| r_i \|$ (voir p. 97), c'est également celui des vecteurs $e_0, \ldots, e_{n-1}$. Les égalités (5.17) admettent

maintenant l'écriture suivante:

$$(f'_{\xi n}, r_i) = -(e_i, r_i) + o(\| r_i \|^2), \quad i = 0, 1, \ldots, n-1.$$

Compte tenu de (5.13), on établit ensuite

$$(f''_{\xi n}\bar{r}_i, r_i) = (f''_{\xi n} r_i, r_i) + (f''_{\xi n}\omega_i, r_i) = (f''_{ic} r_i, r_i) + ((f''_{\xi n} - f''_{ic}) r_i, r_i) + \\ + (f''_{\xi n}\omega_i, r_i) = (e_i, r_i) + o_1(\| r_i \|^2).$$

Donc

$$a_i = \frac{(e_i, r_i) + o(\| r_i \|^2)}{(e_i, r_i) + o_1(\| r_i \|^2)}.$$

En vertu de (5.3)

$$(e_i, r_i) = (f''_{ic} r_i, r_i) \geqslant m \| r_i \|^2, \quad i = 0, 1, \ldots, n-1. \tag{5.18}$$

Par suite, on a, quand $\xi \to \infty$ (i. e. pour $\| r_i \| \to 0$),

$$a_i \to 1, \quad i = 0, 1, \ldots, n-1. \tag{5.19}$$

Puisque $x_{(\xi+1)n} - x_{\xi n} = \sum_{i=0}^{n-1} r_i$, on a

$$x_{(\xi+1)n} - z_{\xi n} = (x_{(\xi+1)n} - x_{\xi n}) - (z_{\xi n} - x_{\xi n}) = \sum_{i=0}^{n-1} (r_i - a_i \bar{r}_i).$$

D'où, compte tenu de (5.13) et (5.19),

$$\| x_{(\xi+1)n} - z_{\xi n} \| = \sum_i o(\| r_i \|),$$

ou, vu (5.8) et (5.9),

$$\| x_{(\xi+1)n} - z_{\xi n} \| = o(\| f'_{\xi n} \|). \tag{5.20}$$

Comme $z_{\xi n} - x_{\xi n} = -(f''_{\xi n})^{-1} f'_{\xi n}$, on a, compte tenu de (5.20),

$$x_{(\xi+1)n} - x_{\xi n} = (x_{(\xi+1)n} - z_{\xi n}) + (z_{\xi n} - x_{\xi n}) = -(f''_{\xi n})^{-1} f'_{\xi n} + \eta_{\xi n},$$

où $\| \eta_{\xi n} \| = o(\| f'_{\xi n} \|)$. Il en résulte l'existence d'une suite de matrices $D^{-1}_{\xi n} \to (f''_{\xi n})^{-1}$ telles que

$$x_{(\xi+1)n} - x_{\xi n} = -D^{-1}_{\xi n} f'_{\xi n} \tag{5.21}$$

(on peut poser par exemple

$$D^{-1}_{\xi n} = (f''_{\xi n})^{-1} + \frac{z_{\xi n} - x_{(\xi+1)n}}{(f'_{\xi n}, f'_{\xi n})} (f'_{\xi n})^* \Big).$$

L'égalité (5.21) révèle la convergence superlinéaire de la suite $\{x_{\xi n}\}$, $\xi = 0, 1, \ldots$; l'estimation correspondante s'obtient de même que dans le théorème 3.1 pour la suite $\{x_k\}$.

Sous l'hypothèse d'avoir (5.6) et les estimations (5.9), (5.10), nous avons donc démontré qu'on a pour $\{x_{\xi n}\}$

$$\| x_{(\xi+1)n} - x_* \| \leqslant \lambda_{\xi n} \| x_{\xi n} - x_* \|, \tag{5.22}$$

où $\lambda_{\xi n}$ tend vers zéro pour $\xi \to \infty$. Mais si l'on a la condition (5.6), l'inégalité (5.22) n'a pas lieu parce que, avec (5.6),

$$\| x_{(\xi+1)n} - x_* \| \geqslant \lambda^n \| x_{\xi n} - x_* \|. \tag{5.23}$$

La contradiction ainsi obtenue signifie que la condition (5.6) (ou (5.23)) ne saurait être satisfaite pour le processus (5.1). Le fait que (5.6) n'est pas satisfaite pour tout k veut dire en fait que dans le processus (5.1) (avec reconstitution) l'inégalité (5.23) n'a pas lieu non plus pour une suite partielle $\{\xi_m\}$, $m = 0, 1, \ldots$ S'il existait une suite $\{\xi_m\}$ telle que

$$\| x_{(\xi_m+1)n} - x_* \| \geqslant \lambda^n \| x_{\xi_m n} - x_* \|, \tag{5.24}$$

on aurait alors dans les méthodes avec reconstitution les estimations (5.9), (5.10) pour tout $\xi_m n \leqslant k \leqslant (\xi_m + 1)\, n$ (nous nous en convaincrons au numéro suivant lorsque nous étudierons les propriétés de tels processus). C'est pourquoi nous constaterions en répétant les mêmes raisonnements qu'aux itérations correspondant à la suite $\{\xi_m\}$ on a l'inégalité (5.22), ce qui contredit (5.24).

Ainsi, dans le cas du processus (5.1) avec reconstitution de la matrice H_k l'inégalité (5.24) ne peut avoir lieu. Il en découle l'existence, pour toute constante $\lambda > 0$, d'un nombre T tel que, lorsque $\xi > T$, la condition (5.22) se trouve satisfaite, i.e. il y a convergence superlinéaire de la suite $\{x_{\xi n}\}$.

3. Etude des propriétés de divers algorithmes. Démontrons les estimations (5.9), (5.10) pour diverses méthodes des directions conjuguées avec reconstitution de la matrice H_k après n pas en supposant que l'inégalité (5.6) (ou (5.8)) est vérifiée.

Quelle que soit la méthode, on doit raisonner par récurrence: on démontre ces estimations pour $i \neq j$, $i, j = 0, 1$; puis, en les supposant justes pour $0 \leqslant i, j \leqslant \tau < n - 1$, on prouve leur validité dans le cas $0 \leqslant i, j \leqslant \tau + 1$.

1. Méthode (4.48). Si la matrice (4.48) se trouve reconstituée au bout d'un nombre fini de pas, la matrice H_k est bornée pour tout k:

$$\| H_k \| \leqslant L, \quad L < \infty. \tag{5.25}$$

Démontrons notre affirmation. En vertu de (5.2),

$$(H_k f'_k, f'_{k+1}) = -(p_k, f'_{k+1}) = 0,$$

donc

$$(H_k e_k, e_k) = (H_k f'_k, f'_k) + (H_k f'_{k+1}, f'_{k+1}). \tag{5.26}$$

La matrice H_k étant définie positive (§ 4), on a $(H_k e_k, e_k) \geqslant$ $\geqslant -(p_k, f'_k) = (p_k, e_k)$, d'où, compte tenu de (5.18),

$$(H_k e_k, e_k) \geqslant \frac{m}{\alpha_k} \| r_k \|^2. \tag{5.27}$$

En se rappelant de plus que $\| e_k \| \leqslant M \| r_k \|$, on obtient de (4.48)

$$\| H_{k+1} \| \leqslant \| H_k \| + \frac{\| r_k \|^2}{m \| r_k \|^2} + \frac{\alpha_k \| H_k \|^2 M^2 \| r_k \|^2}{m \| r_k \|^2}.$$

Compte tenu de la condition (5.4), on s'assure sans peine que $\alpha_{\xi n} \leqslant \bar{\alpha} < \infty$, et l'inégalité de récurrence pour $\| H_{k+1} \|$ entraîne donc l'estimation (5.25) pour $H_{\xi n+1}$. On s'en inspirera pour démontrer par récurrence les inégalités $\alpha_{\xi n+i} \leqslant \bar{\alpha} < \infty$ pour tout $i = 1, \ldots, n-1$. C'est justement en s'appuyant là-dessus qu'on établit (5.25).

Démontrons maintenant que pour $i = 1$ on a les relations

$$(r_1, e_0) = 0, \quad (e_1, r_0) = o(\| r_0 \| \| r_1 \|),$$
$$C_1 \| f_1' \| \leqslant \| r_1 \| \leqslant N_1 \| f_1' \|, \tag{5.28}$$

où les constantes N_1, C_1 sont indépendantes de k et $C_1 > 0$. Les premières estimations (5.28) s'établissent comme suit:

$$(r_1, e_0) = -\alpha_1 (H_1 f_1', e_0) = -\alpha_1 (f_1', H_1 e_0).$$

Mais $H_1 e_0 = r_0$, donc $(r_1, e_0) = -\alpha_1 (f_1', r_0) = 0$. Ensuite

$$(e_1, r_0) = (f_{1c}'' r_1, r_0) = (r_1, f_{0c}'' r_0) + (r_1, (f_{1c}'' - f_{0c}'') r_0) =$$
$$= (r_1, e_0) + o(\| r_1 \| \| r_0 \|) = o(\| r_1 \| \| r_0 \|).$$

Prouvons les estimations pour $\| r_1 \|$. Compte tenu de (5.2) et (5.26), il résulte de (4.48)

$$(H_1 f_1', f_1') = (H_0 f_1', f_1') - \frac{(H_0 e_0, f_1')^2}{(H_0 e_0, e_0)} = (H_0 f_1', f_1') \frac{1}{1 + \frac{(H_0 f_1', f_1')}{(H_0 f_0', f_0')}}.$$

En utilisant (1.14), (1.15) et (5.7) on constate aisément qu'on a pour une fonction vérifiant (5.3)

$$m\left(1 + \frac{m}{M}\right)(f(x) - f_*) \leqslant \| f'(x) \|^2 \leqslant \frac{2M^2}{m}(f(x) - f_*). \tag{5.29}$$

Vu les estimations (5.4) et (5.29), on a sur l'ensemble $S_0 = \{x : f(x) \leqslant f(x_0)\}$:

$$\frac{(H_0 f_1', f_1')}{(H_0 f_0', f_0')} \leqslant \frac{M_0 \| f_1' \|^2}{m_0 \| f_0' \|^2} \leqslant \frac{d_1 (f_1 - f_*)}{d_2 (f_0 - f_*)} \leqslant \frac{d_1}{d_2},$$

avec d_1, d_2 des constantes indépendantes de ξ. Par suite,

$$(H_1 f_1', f_1') \geqslant \frac{(H_0 f_1', f_1')}{1 + \frac{d_1}{d_2}} \geqslant a_1 \| f_1' \|^2, \tag{5.30}$$

où $a_1 = m_0 / \left(1 + \frac{d_1}{d_2}\right)$ ne dépend pas de ξ.

Recourons aux inégalités (5.30) pour évaluer la valeur du paramètre $\alpha_{\xi n+1}$. Comme

$$f_2 - f_1 = \alpha_1 (f_1', p_1) + \frac{\alpha_1^2}{2} (f_1'' p_1, p_1)$$

et α_1 est choisi à partir de la condition (5.2), il est clair que

$$-\frac{(f_1', p_1)}{M \| p_1 \|^2} \leqslant \alpha_1 \leqslant -\frac{(f_1', p_1)}{m \| p_1 \|^2}.$$

Mais en vertu de (5.30) $-(f_1', p_1) = (H_1 f_1', f_1') \geqslant a_1 \| f_1' \|^2$ et, vu (5.25), $\| p_1 \| = \| H_1 f_1' \| \leqslant L \| f_1' \|$; il se trouve, compte tenu de ces estimations, que $\alpha_1 \geqslant \frac{a_1}{ML^2} = \alpha > 0$. Par ailleurs, il découle de (5.30) $\| p_1 \| \geqslant a_1 \| f_1' \|$. On s'en sert et l'on établit sans difficulté que $\alpha_1 \leqslant \frac{L}{ma_1^2} = \bar{\alpha} < \infty$. On a donc

$$N_1 \| f_1' \| = \bar{\alpha} L \| f_1' \| \geqslant \| r_1 \| = \alpha_1 \| H_1^* f_1' \| \geqslant \alpha a_1 \| f_1' \| = C_1 \| f_1' \|,$$

où les constantes N_1, C_1 sont indépendantes de ξ, i.e. on a prouvé les estimations (5.28).

Supposons qu'on a

$$(r_i, e_j) = o(\| r_i \| \| r_j \|), \quad i \neq j, \ 0 \leqslant i, j \leqslant \tau < n-1, \tag{5.31}$$

$$C_i \| f_i' \| \leqslant \| r_i \| \leqslant N_i \| f_i' \|, \quad 0 \leqslant i \leqslant \tau, \tag{5.32}$$

avec N_i, $C_i > 0$ des constantes indépendantes de ξ. Montrons que des estimations analogues ont lieu pour $0 \leqslant i, j \leqslant \tau + 1$:

$$(f_{\tau+1}', r_j) = (f_{j+1}', r_j) + (e_{j+1} + \ldots + e_\tau, r_j), \quad 0 \leqslant j < \tau. \tag{5.33}$$

Etant donné la condition (5.8) et les estimations (5.32), les quantités $\| f_{\tau+1}' \|$, $\| f_i' \|$ et $\| r_i \|$ sont du même ordre infinitésimal pour tous les $0 \leqslant i \leqslant \tau$. Cela permet d'établir moyennant (5.2), (5.31) et (5.33)

$$(f_{\tau+1}', r_j) = o(\| f_{\tau+1}' \| \| r_j \|) = o(\| r_j \|^2), \quad 0 \leqslant j < \tau.$$

Comme $(f_{\tau+1}', r_\tau) = 0$ selon (5.2), on a finalement

$$(f_{\tau+1}', r_j) = o(\| r_j \|^2), \quad 0 \leqslant j \leqslant \tau. \tag{5.34}$$

Evaluons $(H_{\tau+1} f_{\tau+1}', f_{\tau+1}')$. Utilisons la formule (4.48) et prenons en considération (5.26), il vient, quel que soit $0 \leqslant j \leqslant \tau$,

$$(H_{j+1} f_{\tau+1}', f_{\tau+1}') = (H_j f_{\tau+1}', f_{\tau+1}') + \frac{(r_j, f_{\tau+1}')^2}{(r_j, e_j)^2} - \frac{(H_j e_j, f_{\tau+1}')^2}{(H_j e_j, e_j)} \geqslant$$

$$\geqslant \frac{1}{(H_j e_j, e_j)} [(H_j f_{\tau+1}', f_{\tau+1}')(H_j f_{j+1}', f_{j+1}') +$$

$$+ (H_j f_{\tau+1}', f_{\tau+1}')(H_j f_j', f_j') - (H_j f_{j+1}', f_{\tau+1}')^2 - (H_j f_j', f_{\tau+1}')^2 +$$

$$+ 2(H_j f_{j+1}', f_{\tau+1}')(H_j f_j', f_{\tau+1}')].$$

La différence du premier et du troisième terme du numérateur dans le second membre de l'inégalité obtenue est non négative conformément à l'inégalité de Cauchy-Bouniakovski. Vu les estimations (5.34), (5.27), (5.25) et le caractère borné de α_j, $j \leqslant \tau$, on s'assure facilement que le rapport de deux derniers termes du numérateur au dénominateur est une quantité de l'ordre de $o(\| r_j \| \| f'_{\tau+1} \|) = o(\| f'_{\tau+1} \|^2)$. Par suite

$$(H_{j+1}f'_{\tau+1}, f'_{\tau+1}) \geqslant \frac{(H_j f'_{\tau+1}, f'_{\tau+1})(H_j f'_j, f'_j)}{(H_j e_j, e_j)} - o(\| f'_{\tau+1} \|^2).$$

Les estimations (5.32) signifient qu'il existe des constantes a_i indépendantes de ξ telles que $(H_j f'_j, f'_j) = -(p_j, f'_j) \geqslant a_j \| f'_j \|^2$. Compte tenu de la dernière relation et de (5.25)

$$(H_{j+1}f'_{\tau+1}, f'_{\tau+1}) \geqslant \frac{a_j \| f'_j \|^2}{ML \| r_j \|^2} (H_j f'_{\tau+1}, f'_{\tau+1}) - o(\| f'_{\tau+1} \|^2) \geqslant$$

$$\geqslant \tilde{a}_j (H_j f'_{\tau+1}, f'_{\tau+1}) - o(\| f'_{\tau+1} \|^2), \tag{5.35}$$

avec $\tilde{a}_j > 0$ indépendant de ξ (en vertu de (5.32)).

Nous avons indiqué au numéro précédent que $\| f'_k \| \to 0$ pour $k \to \infty$ dans les processus avec reconstitution de H_k. En se rappelant la définie positivité de H_k les inégalités (5.35) entraînent que si, quel que soit ξ, on a $(H_j f'_{\tau+1}, f'_{\tau+1}) \geqslant \gamma_j \| f'_{\tau+1} \|^2$ avec $\gamma_j > 0$ indépendant de ξ, on peut alors trouver une constante $\gamma_{j+1} > 0$ telle que $(H_{j+1}f'_{\tau+1}, f'_{\tau+1}) \geqslant \gamma_{j+1} \| f'_{\tau+1} \|^2$ pour tout ξ. Or, $(H_0 f'_{\tau+1}, f'_{\tau+1}) \geqslant m_0 \| f'_{\tau+1} \|^2$, ce qui permet d'affirmer en évaluant $(H_1 f'_{\tau+1}, f'_{\tau+1})$ qu'il existe une constante γ_1 telle que $(H_1 f'_{\tau+1}, f'_{\tau+1}) \geqslant \gamma_1 \| f'_{\tau+1} \|^2$ pour n'importe quel ξ. Si l'on en tient compte, on peut dire que nos raisonnements par récurrence ci-dessus montrent l'existence d'une constante $a_{\tau+1}$ indépendante de ξ telle que $(H_{\tau+1}, f'_{\tau+1}, f'_{\tau+1}) \geqslant a_{\tau+1} \| f'_{\tau+1} \|^2$. On établit de même que

$$\frac{a_{\tau+1}}{ML^2} \leqslant -\frac{(f'_{\tau+1}, p_{\tau+1})}{M \| p_{\tau+1} \|^2} \leqslant \alpha_{\tau+1} \leqslant -\frac{(f'_{\tau+1}, p_{\tau+1})}{m \| p_{\tau+1} \|^2} \leqslant \frac{L}{m a_{\tau+1}^2}.$$

On a donc

$$N_{\tau+1} \| f'_{\tau+1} \| \geqslant \| r_{\tau+1} \| = \alpha_{\tau+1} \| H_{\tau+1} f'_{\tau+1} \| \geqslant C_{\tau+1} \| f'_{\tau+1} \|. \tag{5.36}$$

Montrons maintenant que

$$H_{\tau+1} e_j = r_j + \eta_j, \quad 0 \leqslant j \leqslant \tau, \tag{5.37}$$

où $\| \eta_j \| = o(\| r_j \|)$.

Multiplions terme à terme (4.48) par e_j, il vient

$$H_{s+1} e_j = H_s e_j + \frac{r_s (r_s, e_j)}{(r_s, e_s)} - \frac{(H_s e_s, e_j) H_s e_s}{(H_s e_s, e_s)}. \tag{5.38}$$

Sous l'hypothèse d'avoir, pour un certain s, $j+1 \leqslant s \leqslant \tau$, les égalités $H_s e_j = r_j + \tilde{\eta}_j$, où $\| \tilde{\eta}_j \| = o(\| r_j \|)$, on s'assure, moyennant les estimations (5.31), (5.27), (5.25) et vu que tous les $\| r_s \|$ sont d'un même ordre infinitésimal, qu'en vertu de (5.38) $H_{s+1} e_j = r_j + \eta_j$, où $\| \eta_j \| = o(\| r_j \|)$. Mais $H_{j+1} e_j = r_j$, et on établit donc par récurrence la validité des égalités (5.37).

Compte tenu de (5.37),

$$(r_{\tau+1}, e_j) = -\alpha_{\tau+1} (H_{\tau+1} f'_{\tau+1}, e_j) = -\alpha_{\tau+1} (f'_{\tau+1}, r_j + \eta_j),$$

par suite, on a, en vertu de (5.34),

$$(r_{\tau+1}, e_j) = o(\| r_j \|^2) + o(\| f'_{\tau+1} \| \| r_j \|), \quad 0 \leqslant j \leqslant \tau.$$

Les inégalités (5.8) et (5.36) montrent que $\| r_{\tau+1} \|$ est du même ordre infinitésimal que $\| f'_{\tau+1} \|$ et donc que $\| r_j \|$, $0 \leqslant j \leqslant \tau$. D'où

$$(r_{\tau+1}, e_j) = o(\| r_{\tau+1} \| \| r_j \|) = o(\| r_{\tau+1} \|^2), \quad 0 \leqslant j \leqslant \tau. \tag{5.39}$$

Cela permet d'obtenir, comme on l'a fait pour $i = 1$,

$$(e_{\tau+1}, r_j) = o(\| r_{\tau+1} \|^2), \quad 0 \leqslant j \leqslant \tau. \tag{5.40}$$

Les relations (5.36), (5.39), (5.40) attestent la validité des estimations (5.31) et (5.32) pour $\tau + 1$ et, par là même, celle des (5.9), (5.10) pour la méthode (4.48) sous l'hypothèse de (5.1) ayant lieu avec reconstitution de la matrice H_k après un nombre fini de pas.

Nos raisonnements se reproduisent intégralement pour le cas où l'on a non pas la condition (5.6) (ou (5.23)), mais l'inégalité (5.24) (ou l'inégalité correspondante $\| f'_{(\xi_{m+1})n} \| \geqslant \delta^n \| f'_{\xi_m n} \|$) et où on ne considère que les itérations correspondant à la sous-suite $\{\xi_m\}$. Les estimations (5.9), (5.10) restent entières pour ces itérations.

Il en résulte (on l'a montré au numéro précédent) la convergence superlinéaire de la méthode étudée.

2. Méthode (4.49). En reconstituant la matrice H_k après un nombre fini de pas, cette matrice est bornée pour tout k. Cela découle de l'inégalité

$$\| H_{k+1} \| \leqslant \| H_k \| + \frac{\| r_k \|^2}{m \| r_k \|^2} + \frac{\| H_k \| M \| r_k \|^2}{m \| r_k \|^2}.$$

Pour $i = 1$

$$(H_1^* f'_1, f'_1) = (H_0 f'_1, f'_1) \geqslant m_0 \| f'_1 \|^2.$$

Si nous tenons compte de ces relations, à force de raisonner comme pour la méthode (4.48), nous établissons les estimations (5.28), puis, en supposant justes (5.31) et (5.32), les estimations (5.34).

Ensuite,

$$(H^*_{\tau+1} f'_{\tau+1}, f'_{\tau+1}) = (H^*_0 f'_{\tau+1}, f'_{\tau+1}) + \sum_{i=0}^{\tau} \frac{(f'_{\tau+1}, r_i)(r_i - H_i e_i, f'_{\tau+1})}{(r_i, e_i)}.$$

D'où, H_k étant bornée et compte tenu de (5.18), (5.34) et du fait que toutes les quantités $\| r_i \|$, $\| e_i \|$, $\| f'_{\tau+1} \|$ sont d'un même ordre infinitésimal,

$$(H^*_{\tau+1} f'_{\tau+1}, f'_{\tau+1}) \geqslant m_0 \| f'_{\tau+1} \|^2 + o(\| f'_{\tau+1} \|^2).$$

Par conséquent, pour $\| f'_{\tau+1} \|$ suffisamment faibles (i.e. pour ξ suffisamment grands), on a

$$(H^*_{\tau+1} f'_{\tau+1}, f'_{\tau+1}) \geqslant a_{\tau+1} \| f'_{\tau+1} \|^2,$$

où $a_{\tau+1} > 0$ est indépendant de ξ. Par suite $\bar{\alpha} \geqslant \alpha_{\tau+1} \geqslant \alpha > 0$ et $C_{\tau+1} \| f'_{\tau+1} \| \leqslant \| r_{\tau+1} \| \leqslant N_{\tau+1} \| f'_{\tau+1} \|$.

Utilisons les égalités

$$H_{s+1} e_i = H_s e_i + (r_s - H_s e_s) \frac{(r_s, e_i)}{(r_s, e_s)}, \quad i+1 \leqslant s \leqslant \tau,$$

tenons compte des estimations (5.31), (5.18) et de la nature bornée de H_k. En raisonnant comme nous l'avons fait pour la méthode (4.48), nous nous assurons que la matrice $H_{\tau+1}$ satisfait aux équations (5.37), ce qui fait qu'on a toujours les estimations (5.39) et (5.40).

La méthode (4.52) se traite de façon analogue.

Remarque. S'il y a $p_i = -H^*_i f'_i = 0$ à une itération du stade initial, on a à recommencer le processus en reconstituant la matrice H_0.

3. Méthode (4.53). En démontrant les résultats relatifs à la matrice H_k, on suit la même voie que pour la méthode (4.49). Notons que la matrice $H_{\tau+1}$ vérifie non plus les conditions (5.37) mais les équations

$$H_{\tau+1} e_j = \eta_j, \quad 0 \leqslant j \leqslant \tau,$$

où $\| \eta_j \| = o(\| r_j \|)$, ce qui simplifie d'ailleurs l'obtention de (5.39).

On procède de même pour la méthode (4.54).

4. Méthode (4.69). La matrice H_k (4.69) définit le vecteur p_k (4.63) avec $\tilde{\beta}_k$ calculé par la formule (4.64).

Pour cette méthode

$$\| H_k \| \leqslant M_0 + \frac{M_0^2 \| f'_k \|^2 + M_0 \| f'_k \| \| H_{k-1} \| \| f'_{k-1} \|}{m_0 \| f'_k \|^2}.$$

En vertu de (5.8), on a, pour tout k, $\frac{\| f'_{k-1} \|}{\| f'_k \|} \leqslant \frac{1}{\delta} = d_0$, où d_0 est indépendant de ξ. Quel que soit k, on a par suite $\| H_k \| \leqslant L$ pour un processus avec reconstitution.

Montrons que, k étant quelconque, on a $\widetilde{\beta}_k \leqslant \overline{\beta} < 1$, i.e.

$$1 - \widetilde{\beta}_k \geqslant 1 - \overline{\beta} = \beta > 0, \tag{5.41}$$

où β ne dépend par de ξ.

Vu (4.65),

$$\widetilde{\beta}_k = \frac{(H_0 f'_k,\ f'_k)}{(H_0 f'_k,\ f'_k) + (1 - \widetilde{\beta}_{k-1})(H_0 f'_{k-1},\ f'_{k-1})} = \frac{1}{1 + (1 - \widetilde{\beta}_{k-1}) \dfrac{(H_0 f'_{k-1},\ f'_{k-1})}{(H_0 f'_k,\ f'_k)}}.$$

En vertu de (5.4) et (5.8) et quel que soit k

$$\frac{(H_0 f'_{k-1},\ f'_{k-1})}{(H_0 f'_k,\ f'_k)} \leqslant \frac{M_0 \| f'_{k-1} \|^2}{m_0 \| f'_k \|^2} \leqslant \gamma < \infty,$$

avec γ indépendant de ξ.

En utilisant (4.65) et (4.64) on établit facilement par récurrence que $0 < \widetilde{\beta}_k < 1$ pour tout k. Donc

$$\widetilde{\beta}_k \leqslant \frac{1}{1 + (1 - \widetilde{\beta}_{k-1})\gamma}.$$

Vu que $\widetilde{\beta}_{\xi n} = 0$, $\xi = 0, 1, \ldots$, on en tire la validité de (5.41). Compte tenu de (5.41), l'égalité (4.65) entraîne pour tout k

$$-(p_k,\ f'_k) \geqslant \beta m_0 \| f'_k \|^2,$$

i.e.

$$\| p_k \| \geqslant \beta m_0 \| f'_k \|.$$

On tient compte de ces inégalités et du caractère borné de H_k pour établir, comme pour les méthodes considérées plus haut, que $0 < \alpha \leqslant \alpha_k \leqslant \overline{\alpha}$ et $C \| f'_k \| \leqslant \| r_k \| \leqslant N \| f'_k \|$.

Les estimations (5.9) se trouvent donc démontrées.

Faisons la même chose en ce qui concerne (5.10). Pour $k = \xi n + 1$

$$(p_1,\ e_0) = -(H_0 f'_1,\ e_0) + \frac{(H_0 f'_1,\ f'_1)}{(H_0 f'_1,\ f'_1) - (p_0,\ f'_0)} [(H_0 f'_1,\ e_0) + (p_0,\ e_0)].$$

Mais $(H_0 f'_1,\ e_0) = (H_0 f'_1,\ f'_1)$ et, conformément à (4.61), $(p_0,\ e_0) = -(p_0,\ f'_0)$, ce qui entraîne $(p_1,\ e_0) = 0$.

Si les estimations (5.31) sont justes, (5.34) se démontrent comme dans la méthode (4.48).

Moyennant (4.63), on établit

$$(p_{\tau+1},\ e_j) = (\widetilde{\beta}_{\tau+1} - 1)(H_0 f'_{\tau+1},\ e_j) + \widetilde{\beta}_{\tau+1}(p_\tau,\ e_j).$$

Evaluons $(H_0 f'_{\tau+1},\ e_j)$. Il résulte de (4.63)

$$H_0 f'_j = \frac{1}{\widetilde{\beta}_j - 1}(p_j - \widetilde{\beta}_j p_{j-1}) \tag{5.42}$$

et on a donc

$$H_0 e_j = \frac{1}{\tilde{\beta}_{j+1} - 1}(p_{j+1} - \tilde{\beta}_{j+1} p_j) - \frac{1}{\tilde{\beta}_j - 1}(p_j - \tilde{\beta}_j p_{j-1}).$$

Utilisons cette expression et les estimations (5.34), (5.41), il vient

$$(f'_{\tau+1}, H_0 e_j) = o(\| p_{\tau+1} \|^2) = o(\| r_{\tau+1} \|^2), \quad 0 \leqslant j < \tau.$$

Puisque c'est également la valeur de (p_τ, e_j) pour $0 \leqslant j < \tau$ (en vertu de (5.31)), il se trouve que

$$(p_{\tau+1}, e_j) = o(\| r_{\tau+1} \|^2), \quad 0 \leqslant j < \tau.$$

Quand $j = \tau$,

$$(p_{\tau+1}, e_\tau) = (\tilde{\beta}_{\tau+1} - 1)(H_0 f'_{\tau+1}, e_\tau) + \tilde{\beta}_{\tau+1}(p_\tau, e_\tau),$$

d'où, compte tenu de (4.64) et (4.61)

$$(p_{\tau+1}, e_\tau) = -\frac{(H_0 f'_{\tau+1}, f'_\tau)(p_\tau, f'_\tau)}{(H_0 f'_{\tau+1}, f'_{\tau+1}) - (p_\tau, f'_\tau)}.$$

Par suite de $(H_0 f'_{\tau+1}, f'_{\tau+1}) > 0$, on a

$$(p_{\tau+1}, e_\tau) \leqslant (H_0 f'_{\tau+1}, f'_\tau).$$

Si on utilise, pour évaluer $(H_0 f'_{\tau+1}, f'_\tau)$, l'expression (5.42), $j = \tau$, on a alors, compte tenu de (5.34) et (5.41),

$$(H_0 f'_{\tau+1}, f'_\tau) = o(\| r_{\tau+1} \|^2)$$

(le fait qu'en vertu de (5.9) toutes les quantités $\| r_i \|$ et $\| f'_i \|$, $i = 0, 1, \ldots, n-1$, sont d'un même ordre infinitésimal y est déjà pris en considération).

Ainsi,

$$(p_{\tau+1}, e_j) = o(\| r_{\tau+1} \|^2), \quad 0 \leqslant j \leqslant \tau,$$

i.e. dans la méthode considérée les estimations (5.39) restent en vigueur. Il en est donc également des (5.40).

Ainsi nous avons établi les estimations (5.9), (5.10) pour (4.69) sous l'hypothèse d'avoir la condition (5.6).

5. Méthode (4.71). En vertu de (4.67) et (5.4),

$$-(p_k, f'_k) \geqslant m_0 \| f'_k \|^2. \tag{5.43}$$

Vu cette inégalité

$$\| H_k \| \leqslant \| H_0 \| + \frac{\| H_0 \| \| f'_k \| \| H^*_{k-1} \| \| f'_{k-1} \|}{m_0 \| f'_{k-1} \|^2}.$$

Le rapport $\| f'_k \| / \| f'_{k-1} \|$ étant borné sur l'ensemble S_0 quel que soit k et $\| H_0 \| \leqslant M_0$, on a

$$\| H_k \| \leqslant M_0 + d \| H_{k-1} \|.$$

où d est une constante. Il s'ensuit qu'en reconstituant la matrice en un nombre fini de pas, H_k reste bornée pour tout k: $\| H_k \| \leqslant L$. Etant donné l'estimation (5.43) on s'assure par suite que $\bar{\alpha} \geqslant \alpha_k \geqslant \geqslant \alpha > 0$ et $N \| f'_k \| \geqslant \| r_k \| \geqslant C \| f'_k \|$ avec un k quelconque.

Les estimations (5.9) sont donc justes pour la méthode (4.71). Prouvons qu'il en est de même pour (5.10):

$$(p_{k+1}, e_k) = -(H_0 f'_{k+1}, e_k) - \frac{(H_0 f'_{k+1}, f'_{k+1})}{(p_k, f'_k)} (p_k, e_k).$$

D'où, compte tenu de (4.61),

$$(p_{k+1}, e_k) = -(H_0 f'_{k+1}, e_k) + (H_0 f'_{k+1}, f'_{k+1}) = (H_0 f'_{k+1}, f'_k). \quad (5.44)$$

Pour $k = \xi n$, (5.44) entraîne

$$(p_1, e_0) = (H_0 f'_1, f'_0) = (f'_1, p_0) = 0.$$

Utilisons (4.66), il vient les expressions suivantes:

$$(p_{\tau+1}, e_j) = -(H_0 f'_{\tau+1}, e_j) + \beta_{\tau+1} (p_\tau, e_j),$$
$$H_0 f'_j = -p_j + \beta_j p_{j-1},$$
$$H_0 e_j = -p_{j+1} + p_j - \beta_{j+1} p_j + \beta_j p_{j-1}.$$

En supposant justes les estimations (5.31), on démontre comme dans la méthode (4.48) les estimations (5.34). Notons également que le coefficient

$$\beta_k = -\frac{(H_0 f'_k, f'_k)}{(p_{k-1}, f'_{k-1})} \leqslant \frac{M_0 \| f'_k \|^2}{m_0 \| f'_{k-1} \|^2} \leqslant \frac{d_1}{d_2}.$$

Etant donné cette expression de β_k et en raisonnant comme pour (4.69), on établit

$$(p_{\tau+1}, e_j) = o(\| r_{\tau+1} \|^2), \quad 0 \leqslant j < \tau,$$

et, de plus,

$$(H_0 f'_\tau, f'_{\tau+1}) = o(\| r_{\tau+1} \|^2),$$

ce qui fait que (5.44) implique

$$(p_{\tau+1}, e_\tau) = o(\| r_{\tau+1} \|^2).$$

Ainsi,

$$(p_{\tau+1}, e_j) = o(\| r_{\tau+1} \|^2), \quad 0 \leqslant j \leqslant \tau,$$

d'où les estimations (5.39). La validité des (5.40) se démontre comme pour (4.48). Pour la méthode considérée on a donc toujours les estimations (5.10).

En étudiant la méthode (4.48) nous avons noté qu'il était possible de démontrer les estimations (5.9), (5.10) en supposant remplie la condition (5.24) — à cette fin on ne répète les raisonnements faits que pour les itérations correspondantes. Cette remarque reste vraie pour les autres méthodes étudiées, ce qui atteste leur convergence superlinéaire.

4. Etude de la vitesse de convergence.

1. Supposons que la matrice $f''(x)$ vérifie, en plus des conditions (5.3), la condition de Lipschitz (2.8). La vitesse de convergence de la suite $\{x_{\xi n}\}$ se prête alors à une évaluation plus précise.

Dans le souci de faciliter les références nous écrirons différentes relations valables sous la condition (5.3) (nombre d'entre elles nous ont déjà servi à plusieurs reprises):

$$m \| x - x_* \| \leqslant \| f'(x) \| \leqslant M \| x - x_* \|; \tag{5.45}$$

$$d_1 [f(x) - f(x_*)] \leqslant \| f'(x) \|^2 \leqslant d_2 [f(x) - f(x_*)] \tag{5.46}$$

(les constantes d_1, d_2 sont indépendantes du choix du point x);

$$m \| r_k \| \leqslant \| e_k \| \leqslant M \| r_k \|. \tag{5.47}$$

Soient x, y deux points quelconques et

$$f(y) \leqslant f(x). \tag{5.48}$$

On établit alors, compte tenu de (5.46), que

$$\| f'(y) \| \leqslant C \| f'(x) \|. \tag{5.49}$$

C désigne ici (et partout dans ce numéro) des constantes (non nulles) indépendantes du choix des points x, $y \in E^n$.

On a sous (5.48), en vertu de (5.45) et (5.49),

$$\| y - x_* \| \leqslant C \| x - x_* \|. \tag{5.50}$$

2. Supposons que les processus itératifs étudiés vérifient l'estimation

$$\| f'_{\xi n+i} \| \| f'_{\xi n+i+1} \| \leqslant \lambda_\xi \| f'_{\xi n+j} \|, \qquad 0 \leqslant i < j \leqslant n-1. \tag{5.51}$$

λ_ξ désignera désormais diverses variables tendant vers zéro quand $\xi \to \infty$.

Dans la suite nous nous bornerons à étudier les propriétés de la méthode (4.48). Les résultats obtenus (lemme 5.1, théorème 5.2) seront cependant valables pour d'autres algorithmes de directions conjuguées.

LEMME 5.1. *Si une fonction deux fois continûment dérivable $f(x)$ satisfaisant aux conditions* (5.3) *et* (2.8) *est minimisée par le processus* (5.1) *où la matrice H_k est construite par la formule* (4.48) *et si les inégalités* (5.51) *ont lieu, on a alors les estimations* (5.9) *et*

$$| (r_{\xi n+i}, e_{\xi n+j}) | \leqslant C \| r_{\xi n+t} \| \| r_{\xi n+t+1} \|, \tag{5.52}$$

$$t = \min \{i, j\}, \quad i \neq j, \quad 0 \leqslant i, \quad j \leqslant n-1.$$

La démonstration est analogue à celle des estimations (5.9) et (5.10) pour la méthode (4.48) (n° 3); seul l'ordre infinitésimal de certaines quantités est précisé. On se passe donc de la démonstration même pour ne s'attarder que sur les modifications survenues.

Pour $i = 1$

$$(e_1, r_0) = (r_1, e_0) + (r_1, (f''_{1c} - f''_{0c}) r_0).$$

Compte tenu de (2.8), on obtient

$$\| f''_{1c} - f''_{0c} \| \equiv \| f''(x_1 + \theta_1 r_1) - f''(x_0 + \theta_0 r_0) \| \leqslant$$
$$\leqslant R (\| x_1 - x_0 \| + \| r_1 \| + \| r_0 \|) \leqslant R (2 \| r_0 \| + \| r_1 \|).$$

Par suite de (5.50) et (5.45)

$$\| r_k \| \leqslant \| x_k - x_* \| + \| x_{k+1} - x_* \| \leqslant C \| f'_k \|. \quad (5.53)$$

Moyennant (5.53) et (5.49) on établit $\| r_1 \| \leqslant C \| f'_1 \| \leqslant C \| f'_0 \|$. Vu que $\| r_0 \| \geqslant C \| f'_0 \|$, on trouve $\| r_1 \| \leqslant C \| r_0 \|$. Donc $\| f''_{1c} - f''_{0c} \| \leqslant C \| r_0 \|$, et on a, compte tenu de cette inégalité,

$$| (e_1, r_0) | \leqslant C \| r_0 \|^2 \| r_1 \|.$$

Supposons les estimations (5.9) et (5.52) vérifiées pour $0 \leqslant i, j \leqslant \leqslant \tau < n - 1$. Puisque $(f'_{j+1}, r_j) = 0$, on a alors

$$| (f'_{\tau+1}, r_j) | = | (e_\tau + \ldots + e_{j+1}, r_j) | \leqslant C \| r_j \|^2 \| r_{j+1} \|,$$
$$0 \leqslant j < \tau, \quad (5.54)$$

d'où, compte tenu de (5.53),

$$| (f'_{\tau+1}, r_j) | \leqslant C \| r_j \| \| f'_j \| \| f'_{j+1} \|. \quad (5.55)$$

Si (5.51) a lieu, $\| f'_j \| \| f'_{j+1} \| \leqslant \lambda_\xi \| f'_{\tau+1} \|$. Il résulte donc de (5.55)

$$| (f'_{\tau+1}, r_j) | \leqslant \lambda_\xi \| r_j \| \| f'_{\tau+1} \|, \ 0 \leqslant j < \tau.$$

Comme $(f'_{\tau+1}, r_\tau) = 0$, on a finalement

$$| (f'_{\tau+1}, r_j) | \leqslant \lambda_\xi \| r_j \| \| f'_{\tau+1} \|, \quad 0 \leqslant j \leqslant \tau. \quad (5.56)$$

Utilisons les estimations (5.27), (5.56) et (5.25), il vient également

$$\frac{(H_j f'_j, f'_{\tau+1})^2}{(H_j e_j, e_j)} \leqslant \frac{(f'_{\tau+1}, r_j)^2}{m \| r_j \|^2} \leqslant \lambda_\xi \| f'_{\tau+1} \|^2, \ 0 \leqslant j \leqslant \tau, \quad (5.57)$$

$$\frac{| (H_j f'_{j+1}, f'_{\tau+1}) (H_j f'_j, f'_{\tau+1}) |}{(H_j e_j, e_j)} \leqslant \frac{C \| f'_{j+1} \| \| f'_{\tau+1} \| | (f'_{\tau+1}, r_j) |}{m \| r_j \|^2} \leqslant$$
$$\leqslant \frac{\lambda_\xi \| f'_{\tau+1} \|^2 \| f'_{j+1} \|}{\| r_j \|}. \quad (5.58)$$

Etant donné $\| r_j \| \geqslant C \| f'_j \|$, $0 \leqslant j \leqslant \tau$, et vu (5.53), (5.49), on a

$$\| r_i \| \leqslant C \| f'_i \| \leqslant C \| f'_j \| \leqslant C \| r_j \|, 0 \leqslant j \leqslant i \leqslant \tau. \quad (5.59)$$

Compte tenu de (5.59),

$$\frac{\lambda_\xi \| f'_{\tau+1} \|^2 \| f'_{j+1} \|}{\| r_j \|} \leqslant \lambda_\xi \| f'_{\tau+1} \|^2, \quad 0 \leqslant j \leqslant \tau. \quad (5.60)$$

En utilisant (5.57), (5.58) et (5.60), on établit de même que dans le numéro 3 que $(H_{\tau+1} f'_{\tau+1}, f'_{\tau+1}) \geqslant a_{\tau+1} \| f'_{\tau+1} \|^2$.

Montrons maintenant que

$$H_{\tau+1}e_j = r_j + \eta_{j,\tau+1},\ 0 \leqslant j \leqslant \tau, \tag{5.61}$$

où

$$\|\eta_{j,\tau+1}\| \leqslant C \sum_{\nu=j+1}^{\tau} \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_\nu\|}, \qquad 0 \leqslant j < \tau \quad \|\eta_{\tau,\tau+1}\| = 0. \tag{5.62}$$

En effet,

$$H_{s+1}e_j = H_s e_j + \frac{r_s(r_s, e_j)}{(r_s, e_s)} - \frac{(e_s, H_s e_j) H_s e_s}{(H_s e_s, e_s)}. \tag{5.63}$$

On a $H_{j+1}e_j = r_j$.

Utilisons les estimations (5.52) (justes par hypothèse pour $0 \leqslant s$, $j \leqslant \tau$), (5.25), (5.27), (5.47) et le fait que $|\alpha_i| \leqslant C$, $0 \leqslant i \leqslant \tau$, il vient pour $s = j + 1$

$$\frac{\|(e_{j+1}, H_{j+1}e_j) H_{j+1}e_{j+1}\|}{(H_{j+1}e_{j+1}, e_{j+1})} = \frac{\|(e_{j+1}, r_j) H_{j+1}e_{j+1}\|}{(H_{j+1}e_{j+1}, e_{j+1})} \leqslant C \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_{j+1}\|}.$$

Etant donné que $(r_{j+1}, e_j) = 0$, on obtient à partir de (5.63)

$$H_{j+2}e_j = r_j + \eta_{j,j+2},\ \|\eta_{j,j+2}\| \leqslant C \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_{j+1}\|}.$$

Supposons que pour un $j + 1 < s \leqslant \tau$ on a

$$H_s e_j = r_j + \eta_{j,s}, \quad \|\eta_{j,s}\| \leqslant C \sum_{\nu=j+1}^{s-1} \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_\nu\|}.$$

En vertu des mêmes conditions que pour $s = j + 1$ on a alors

$$\frac{\|(e_s, H_s e_j) H_s e_s\|}{(H_s e_s, e_s)} \leqslant C \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_s\|} + C \sum_{\nu=j+1}^{s-1} \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_\nu\|},$$

$$\frac{\|r_s(r_s, e_j)\|}{(r_s, e_s)} \leqslant C \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_s\|}.$$

On établit en portant ces estimations dans (5.63)

$$H_{s+1}e_j = r_j + \eta_{j,s+1},\ \|\eta_{j,s+1}\| \leqslant C \sum_{\nu=j+1}^{s} \frac{\|r_j\|^2 \|r_{j+1}\|}{\|r_\nu\|}.$$

Ainsi, (5.61) est valide par récurrence.

Démontrons maintenant (5.52) quand $i = \tau + 1$. Compte tenu de (5.61),

$$(r_{\tau+1}, e_j) = -\alpha_{\tau+1}(H_{\tau+1}f'_{\tau+1}, e_j) = -\alpha_{\tau+1}(f'_{\tau+1}, r_j + \eta_{j,\tau+1}). \tag{5.64}$$

Par suite de (5.59) et (5.49)

$$\|r_\nu\| \geqslant C\|f'_\tau\| \geqslant C\|f'_{\tau+1}\|,\ 0 \leqslant \nu \leqslant \tau. \tag{5.65}$$

Utilisons cette inégalité et (5.62), il vient $|(f'_{\tau+1}, \eta_{j,\tau+1})| \leqslant \leqslant C \| r_j \|^2 \| r_{j+1} \|$. Portons celle-ci et les estimations (5.54) dans (5.64) et prenons en considération $|\alpha_{\tau+1}| \leqslant C$, nous établissons

$$|(r_{\tau+1}, e_j)| \leqslant C \| r_j \|^2 \| r_{j+1} \|, \ 0 \leqslant j \leqslant \tau. \tag{5.66}$$

Ensuite,

$$(e_{\tau+1}, r_j) = (r_{\tau+1}, e_j) + (r_{\tau+1}, (f''(x_{\tau+1} + \theta_{\tau+1} r_{\tau+1}) - f''(x_j + \theta_j r_j)) r_j). \tag{5.67}$$

Etant donné (5.65) et (5.53), on a $\| r_{\tau+1} \| \leqslant C \| r_j \|$, $0 \leqslant j \leqslant \tau$. Compte tenu de (2.8) on a donc

$$\| f''(x_{\tau+1} + \theta_{\tau+1} r_{\tau+1}) - f''(x_j + \theta_j r_j) \| \leqslant R (\| x_{\tau+1} - x_j \| + \\ + \| r_{\tau+1} \| + \| r_j \|) \leqslant C \| r_j \|.$$

Reprenons (5.67) avec cette estimation et (5.66), nous trouvons

$$|(e_{\tau+1}, r_j)| \leqslant C \| r_j \|^2 \| r_{j+1} \|, \ 0 \leqslant j \leqslant \tau.$$

Ainsi, nous avons établi par récurrence les estimations (5.9) et (5.52) et achevé la démonstration du lemme.

THÉORÈME 5.2. *Etant donné une fonction $f(x)$ deux fois continûment dérivable et une matrice $f''(x)$ vérifiant les conditions* (5.3) *et* (2.8), *si l'on recherche le minimum de $f(x)$ par l'algorithme* ((5.1), (4.48)) *pour tous les ξ suffisamment grands on a l'estimation*

$$\| x_{(\xi+1)n} - x_* \| \leqslant C \| x_{\xi n+1} - x_* \| \| x_{\xi n} - x_* \|. \tag{5.66}$$

DÉMONSTRATION. Par suite de (5.45), l'estimation (5.66) est équivalente à

$$\| f'_n \| \leqslant C \| f'_1 \| \| f'_0 \|. \tag{5.67}$$

Supposons que celle-ci n'a pas lieu pour tous les ξ suffisamment grands. Il existe alors une suite partielle infinie $\{\xi_m\}$ telle que les points correspondants vérifient les inégalités

$$\| f'_0 \| \| f'_1 \| \leqslant \lambda_{\xi_m} \| f'_n \|, \ \lambda_{\xi_m} \to 0 \text{ pour } \xi_m \to \infty. \tag{5.68}$$

On estime sans restreindre la généralité que la suite partielle $\{\xi_m\}$ coïncide avec toute la suite $\xi = 0, 1, \ldots$ On s'assure en tenant compte de (5.49) que les estimations (5.51) ont lieu dès qu'on a (5.68). Si ces dernières sont donc supposées justes, les exigences du théorème 5.2 garantissent les conditions du lemme 5.1. Ainsi, la validité de (5.68) entraîne celle de (5.9) et (5.52). Donc

$$|(f'_n, r_j)| = |(e_{n-1} + \ldots + e_{j+1}, r_j)| \leqslant C \| r_j \|^2 \| r_{j+1} \|, \\ 0 \leqslant j \leqslant n-2.$$

On montre comme pour les inégalités (5.56) que

$$|(f'_n, r_j)| \leqslant \lambda_\xi \|f'_n\| \|r_j\|, \ \lambda_\xi \to 0 \text{ pour } \xi \to \infty, \qquad (5.69)$$
$$0 \leqslant j \leqslant n-1.$$

Montrons que dans le cas de l'estimation (5.51) le système $r_0, \ldots, r_{n-1}$ est linéairement indépendant. Commençons par dire que (5.51) implique, compte tenu des estimations (5.9),

$$\|r_i\| \|r_{i+1}\| \leqslant \lambda_\xi \|r_j\|, \ 0 \leqslant \tau < j \leqslant n-1. \qquad (5.70)$$

Vu (5.70) et (5.47), les estimations (5.52) peuvent prendre la forme

$$|(r_i, e_j)| \leqslant \lambda_\xi \|r_i\| \|r_j\| \leqslant \lambda_\xi \|r_i\| \|e_j\|, \ \lambda_\xi \to 0,$$
$$i \neq j, \ i, j = 0, 1, \ldots, n-1. \qquad (5.71)$$

Introduisons la notation $\bar{r}_i = r_i/\|r_i\|$ et soit

$$\|\varphi\| = \min_{\sum_{i=0}^{n-1} |\beta_i| = 1} \left\|\sum_{i=0}^{n-1} \beta_i \bar{r}_i\right\| = \left\|\sum_{i=0}^{n-1} \bar{\beta}_i \bar{r}_i\right\|.$$

On a alors

$$|(\varphi, e_j)| \geqslant \left| |\bar{\beta}_j (r_j, e_j)| - \left|\sum_{i \neq j} \bar{\beta}_i (\bar{r}_i, e_j)\right|\right|. \qquad (5.72)$$

Puisque $\sum_{i=0}^{n-1} |\bar{\beta}_i| = 1$, on a $|\bar{\beta}_j| \geqslant \beta > 0$ pour au moins l'un des indices $j \in 0, n-1$. Ceci étant, on a, compte tenu de (5.18) et (5.47),

$$|\bar{\beta}_j (r_j, e_j)| \geqslant C \|r_j\| \geqslant C \|e_j\|.$$

Quand $i \neq j$, on a, vu (5.71),

$$|\bar{\beta}_i (\bar{r}_i, e_j)| \leqslant \lambda_\xi |\bar{\beta}_i| \|\bar{r}_i\| \|e_j\| = \lambda_\xi |\bar{\beta}_i| \|e_j\|,$$
$$\lambda_\xi \to 0 \text{ pour } \xi \to \infty.$$

Récrivant (5.72) moyennant les inégalités obtenues, on s'assure que pour ξ suffisamment grands on a $|(\varphi, e_j)| \geqslant C \|e_j\|$, i.e.

$$\|\varphi\| \geqslant C. \qquad (5.73)$$

D'où l'indépendance linéaire du système $r_0, \ldots, r_{n-1}$. On démontre de plus sans peine, moyennant (5.73), l'affirmation suivante : si $\psi_0, \ldots, \psi_{n-1}$ est le système dual de $r_0, \ldots, r_{n-1}$, on a pour ξ suffisamment grands

$$\|r_i\| \|\psi_i\| \leqslant C, \ 0 \leqslant i \leqslant n-1. \qquad (5.74)$$

On démontre enfin que dans les conditions (5.69) et (5.74) le système de vecteurs f'_n, r_0, . . ., r_{n-1} est lui aussi linéairement indépendant. Supposons en effet que

$$f'_n = \sum_{i=0}^{n-1} \gamma_i \psi_i = \sum_{i=0}^{n-1} (f'_n, r_i) \psi_i.$$

Ceci étant, on a en vertu de (5.69) et (5.74)

$$\| f'_n \| \leqslant C\lambda_\xi \| f'_n \|.$$

Du moment que $\lambda_\xi \to 0$, cette dernière inégalité n'a pas lieu pour ξ suffisamment grands. D'où l'indépendance linéaire du système f'_0, r_0, . . ., r_{n-1}.

Sous l'hypothèse de l'absence de (5.66) pour tous les $\xi \geqslant \xi_0$ (avec ξ_0 un nombre suffisamment grand) nous avons démontré l'indépendance linéaire dans E^n du système de $n+1$ vecteurs f'_n, r_0, ..., r_{n-1}, ce qui est impossible. Ainsi, l'hypothèse de départ est fausse, i.e. on a bien l'estimation (5.66).

Le théorème est démontré.

5. Discussion des résultats. Ainsi, nous avons établi l'applicabilité de toutes les techniques du § 4 dans la recherche du minimum de fonctions non quadratiques, la convergence pouvant être garantie pour une classe de fonctions à minimiser par les méthodes du gradient. Quand les méthodes des directions conjuguées s'appliquent à des fonctions fortement convexes, leur convergence est au moins superlinéaire.

En évaluant la vitesse de convergence des méthodes des directions conjuguées nous avons agi un peu autrement que pour les méthodes d'autres classes des paragraphes précédents: nous avons considéré, au lieu de $\{x_k\}$, la suite $\{x_{\xi n}\}$, i.e. nous avons en fait pris pour itération la réunion de n itérations ordinaires du processus $x_{\xi n}, x_{\xi n+1}, \ldots$ . . ., $x_{\xi n+n-1}$. De tels processus peuvent en général converger plus lentement que les méthodes des directions duales (§ 3) et à plus forte raison que la méthode de Newton (§ 2) (i.e. la décroissance de la fonction par itération $| f_{k+1} - f_k |$ peut s'avérer moindre pour la classe considérée que pour les méthodes des § 2,3 et le rapport $\| x_{k+1} - x_* \| / \| x_k - x_* \|$ plus important). Si, par exemple, dans un algorithme

$$x_{k+1} - x_k = -D_k^{-1} f'_k, \tag{5.75}$$

et dans une méthode des directions conjuguées

$$x_{(\xi+1)n} - x_{\xi n} = -D_{\xi n}^{-1} f'_{\xi n},$$

et $D_k = D_{\xi n} \to f''_k$, cela signifie que n itérations de la dernière technique sont équivalentes au sens de la convergence à une itération

de (5.75). N'empêche que du point de vue pratique la vitesse de convergence des méthodes de la classe considérée est sensiblement supérieure à celle des méthodes du gradient.

D'autre part, les premières coûtent par itération (on l'a déjà signalé au § 4) à peu près la même chose que les deuxièmes.

Nous sommes donc autorisés à affirmer que lorsqu'il s'agit d'un problème de minimisation, les méthodes des directions conjuguées se révèlent parmi les plus efficaces.

Dans ce paragraphe nous nous sommes bornés à plusieurs algorithmes construits au § 4. On pourrait très bien étudier les propriétés d'autres algorithmes de ce type construits moyennant le schéma général du § 4 et la technique en serait sensiblement la même que celle du § 5. En effet, la seule différence dans la démonstration du théorème 5.1 est d'employer des procédés légèrement différents pour analyser les propriétés de la matrice H_k. Mais. quelle que soit la méthode de la classe en question, les vecteurs u_k et v_k servant à construire H_{k+1} ne sont que diverses combinaisons linéaires des vecteurs r_k et $H_k^* e_k$ (voir (4.32)). Or, en choisissant les algorithmes des § 4,5 nous avons procédé de sorte que les matrices H_k soient construites en utilisant des combinaisons variées de ces éléments.

Servons-nous des résultats obtenus pour comparer les propriétés de divers algorithmes dans la recherche du minimum de fonctions non quadratiques.

Les résultats du théorème 5.2 (estimation (5.66)) montrent que la vitesse de convergence de la suite $\{x_{\xi n}\}$ dépend essentiellement des propriétés de la matrice $H_{\xi n}$. Si pour $\xi \to \infty$

$$H_{\xi n} \to (f''_{\xi n})^{-1}, \tag{5.76}$$

alors

$$\frac{\| x_{\xi n+1} - x_* \|}{\| x_{\xi n} - x_* \|} \to 0$$

et la convergence devient plus rapide. Dans la pratique, ce fait présente un intérêt particulier pour les algorithmes dans lesquels

$$H_n = A^{-1} \tag{5.77}$$

dans la minimisation d'une fonction quadratique. En font partie les algorithmes (4.48), (4.49), (4.52). Lorsqu'on aborde un problème par l'un de ces algorithmes et qu'on est dans la condition (5.76), les considérations ci-dessus nous font nous abstenir de reconstituer la matrice H_k.

Dans la méthode (4.70), la condition (5.77) n'a pas lieu et sa variante sans reconstitution ne possède donc pas d'avantage (au sens de la convergence) sur la mise en œuvre avec reconstitution. C'est également le cas d'algorithmes tels que dans la minimisation d'une fonction quadratique ou bien $H_n = H_0$ (p. ex. (4.69), (4.71)) ou

bien H_n s'approche de H_0 (c'est la matrice H_n de la méthode (4.70), qui agit sur le système de vecteurs linéairement indépendants $e_0, \ldots, e_{n-1}$ (excepté ce dernier) de même que H_0). Ce n'est donc pas la peine de considérer les versions sans reconstitution de H_k.

La convergence des méthodes (4.70), (4.71) devient cependant plus rapide si l'on utilise, au lieu d'une matrice H_0 fixe, une suite de matrices $H_{\xi 0}$ définies positives telles que

$$H_{\xi 0} \to (f''_{\xi n})^{-1}. \tag{5.78}$$

Si l'on se place dans les conditions du lemme 5.1, on construit les matrices $H_{\xi 0}$ vérifiant la condition (5.78) par recours aux vecteurs $r_{\xi n}, r_{\xi n+1}, \ldots, r_{\xi n+n-1}$ et la formule utilisée est

$$H_{(\xi+1)0} = \sum_{i=0}^{n-1} \frac{r_{\xi n+i} r^*_{\xi n+i}}{(r_{\xi n+i}, e_{\xi n+i})}.$$

A la lumière de ces considérations, de toutes les méthodes des directions conjuguées, ce sont celles jouissant de la propriété (5.77) qui possèdent (au sens de la convergence) la plus bonne efficacité dans la minimisation de fonctions strictement convexes.

Dans la pratique, il y a certes exceptions à cette règle en ce sens, par exemple, qu'avec la méthode (4.70) on trouve la solution d'un problème de minimisation (avec une précision déterminée) au bout d'un nombre moindre d'itérations que, disons, moyennant la méthode (4.48). Le fait est (nous y avons insisté à plusieurs reprises) que la vitesse de convergence d'une méthode concrète subit l'influence de nombreux facteurs complémentaires comme les erreurs de calcul, un choix inexact de la valeur de α_k et d'autres encore et que la sensibilité aux perturbations varie d'un procédé à l'autre. Une comparaison de la vitesse de convergence n'a en outre de sens que dans un voisinage suffisamment restreint du point de minimum, toute confrontation de la puissance de divers algorithmes dans une région éloignée de ce point n'étant possible que par la voie d'expériences numériques.

Plusieurs auteurs (J. D. Pearson [92], J. Greenstadt [69], B. Poliak [28; 2], H. Y. Huang, A. V. Levy [76]) citent les résultats de la résolution numérique de problèmes par les méthodes des directions conjuguées. Le dernier ouvrage fournit la plus complète analyse comparative de divers algorithmes. Dans l'ensemble les résultats des expériences sur machine confirment la puissance des méthodes vérifiant la condition (5.77). D'autre part, la méthode (4.71) s'avère la plus efficace (par comparaison avec le processus sans reconstitution) si la matrice est reconstituée en n itérations. Visiblement, la technique (4.70) est à réaliser elle aussi avec reconstitution de H_k.

Pour conclure, arrêtons-nous sur des questions relatives à la longueur du pas dans les méthodes de la classe considérée. On sait déjà que dans les méthodes des directions conjuguées le choix du

pas se fait à partir de la condition de minimum de la fonction dans la direction de déplacement. Nous avons signalé plus d'une fois l'inconvénient de cette procédure, à savoir la nécessité d'effectuer de nombreux calculs de la fonction, ce qui en détermine un coût assez élevé dans des problèmes où le calcul de la fonction est laborieux. Dans certains cas, ce procédé de choix du pas exclut en général toute résolution si, par exemple, la valeur du paramètre α_k varie fortement à chaque pas. Ce défaut des méthodes en question est mis à l'index dans de nombreux travaux (p. ex. C. G. Broyden [46 ; 2], B .Pchénitchny [29 ; 3], W. C. Davidon [54 ; 2], M. J. D. Powell [94; 1], R. Fletcher [60 ; 1]). Pour en venir à bout, ces auteurs considèrent des méthodes où le choix de α_k ne fait que garantir un certain degré de décroissance de la fonction. Quant au reste, ils s'inspirent, en construisant les méthodes, des mêmes idées que celles exposées plus haut (l'ouvrage de B. Pchénitchny en est une exception).

Il est infiniment plus difficile d'étudier les propriétés de méthodes où le choix du pas n'est plus lié à la recherche du minimum de la fonction dans la direction de déplacement si bien que nombre d'entre elles n'ont pas reçu de justification théorique même lorsque la fonction à minimiser est quadratique.

Du point de vue du choix de la valeur de α_k, on constate l'avantage des méthodes des directions duales du § 3 dont l'évaluation de la convergence est plus précise. Mais elles exigent plus de mémoires (il faut stocker deux matrices $n \times n$; voir § 3) et permettent donc de résoudre des problèmes de minimisation en dimension moindre. Il est vrai qu'on peut réduire l'occupation de mémoire nécessaire en choisissant les vecteurs r_k coïncidant avec les axes de coordonnées, mais avec cet artifice on a à calculer la dérivée deux fois par itération, d'où un coût plus grand de tels algorithmes.

§ 6. Méthodes directes

1. Préliminaires. Les méthodes de minimisation que nous avons considérées nécessitent à chaque itération le calcul de la fonction $f(x)$, de son gradient $f'(x)$ (§§ 1, 3, 4, 5) et, dans la méthode de Newton (§ 2), de la matrice des dérivées secondes $f''(x)$. Nous avons insisté sur le caractère laborieux du calcul de $f''(x)$, qui constitue, dans de nombreux cas, l'élément le plus coûteux et ardu du processus itératif, et nous avons élaboré les méthodes des §§ 3-5 dans le but de l'éviter. Dans certains problèmes la recherche du gradient peut s'avérer elle aussi notablement plus compliquée que celle de la fonction (il se peut même qu'il soit impossible d'exprimer $f'(x)$ analytiquement). La résolution du problème se fera alors de préférence par des techniques n'exigeant que le calcul des valeurs de la fonction.

La recherche du gradient par une formule analytique est remplacée par un calcul approché en utilisant par exemple l'approximation aux différences de dérivées partielles. On construit ainsi des variantes des méthodes des paragraphes précédents, qui se bornent au calcul de la fonction. A force d'exiger une approximation d'une précision déterminée et d'imposer certaines contraintes supplémentaires sur la construction du processus itératif, on obtient des méthodes modifiées dont les propriétés (convergence et sa vitesse) se rapprochent de celles des algorithmes de départ utilisant le calcul de $f'(x)$, $f''(x)$ par des expressions analytiques.

L'étude de méthodes directes rend un autre service : en établissant la précision de l'approximation de dérivées pour laquelle les propriétés de tels algorithmes coïncident avec celles des méthodes initiales correspondantes, on définit en fait les erreurs de calcul permises qui n'altèrent pas les propriétés des algorithmes (avec calcul de $f'(x)$, $f''(x)$).

Dans le présent paragraphe nous n'étudierons que des algorithmes construits dans le cadre des méthodes des directions duales du § 3, si bien que cette appellation sera conservée. Nous nous arrêterons également sur des algorithmes qui réalisent l'idée des directions conjuguées sans le calcul du gradient ou son approximation aux différences.

2. Construction des méthodes des directions duales. Dans ces méthodes, les approximations successives sont générées par la formule

$$x_{k+1} = x_k - \alpha_k D_k^{-1} g_k, \tag{6.1}$$

où D_k est une matrice $n \times n$ et g_k un vecteur. A la différence des méthodes ci-dessus le facteur scalaire α_k peut prendre des valeurs positives ou négatives selon la direction de descente ($-D_k^{-1}g_k$ ou $D_k^{-1}g_k$) de la fonction $f(x)$. On peut encore estimer que $\alpha_k \geqslant 0$ et prendre alors pour direction de déplacement soit le vecteur $p_k = -D_k^{-1}g_k$, soit poser $p_k = D_k^{-1}\, g_k$ de façon qu'on ait

$$(f'_k,\ p_k) < 0. \tag{6.2}$$

Comme au § 3 on considère que $f(x)$ est fortement convexe et admet des dérivées secondes continues.

Construction de la matrice D_k et du vecteur g_k. Définissons les vecteurs

$$\vartheta_k = \left(\frac{f(x_k+\mu_k v_1)-f(x_k)}{\mu_k}, \ldots, \frac{f(x_k+\mu_k v_n)-f(x_k)}{\mu_k}\right),$$

$$\varphi_k = \left(\frac{f(y_k+\mu_k v_1)-f(y_k)}{\mu_k}, \ldots, \frac{f(y_k+\mu_k v_n)-f(y_k)}{\mu_k}\right),$$

$$\psi_k = \varphi_k - \vartheta_k,$$

où $0 < |\mu_k| \leqslant \|r_k\|^t$, $t > 1$, y_k, r_k sont les termes de la suite (3.5) et v_i le vecteur unité de l'axe correspondant.

LEMME 6.1. *Etant donné une suite bornée* $\{x_k\}$, $\|x_{k+1} - x_k\| \to 0$ *pour* $k \to \infty$ *et une matrice* D_k *définie, quel que soit* $k \geqslant n-1$, *par le système d'équations*

$$D_k r_{k-i} = \psi_{k-i}, \quad i = 0, 1, \ldots, n-1, \tag{6.3}$$

où r_{k-i} *sont les termes de la suite* (3.5), *on a*

$$\lim_{k\to\infty} \|D_k - f''_k\| = 0.$$

La DÉMONSTRATION coïncide pour l'essentiel avec celle du lemme 3.1, aussi seules les différences nous intéresseront.

Les composantes des vecteurs ϑ_k et φ_k peuvent se mettre sous la forme

$$\vartheta_k^j = \left.\frac{\partial f}{\partial x^j}\right|_{x=x_k+\theta_j\mu_k v_j}, \qquad 0 \leqslant \theta_j \leqslant 1,$$

$$\varphi_k^j = \left.\frac{\partial f}{\partial x^j}\right|_{x=y_k+\xi_j\mu_k v_j}, \qquad 0 \leqslant \xi_j \leqslant 1,$$

$$j = 1, \ldots, n.$$

Vu ces formules et la continuité des dérivées secondes on s'assure aisément de la justesse des estimations

$$\|\vartheta_k - f'(x_k)\| \leqslant C_1 |\mu_k| n^{1/2} \leqslant C_2 \|r_k\|^t, \tag{6.4}$$

$$\|\varphi_k - f'(y_k)\| \leqslant C_3 \|\mu_k\| n^{1/2} \leqslant C_4 \|r_k\|^t, \tag{6.5}$$

avec C_2, $C_4 < \infty$.

Ecrivons le vecteur ψ_{k-i} comme suit:

$$\psi_{k-i} = f'(y_{k-i}) - f'(x_{k-i}) + (\varphi_{k-i} - f'(y_{k-i})) - (\vartheta_{k-i} - f'(x_{k-i})).$$

En conservant la notation $e_{k-i} = f'(y_{k-i}) - f'(x_{k-i})$, on obtient

$$D_k r_{k-i} = e_{k-i} + (\varphi_{k-i} - f'(y_{k-i})) - (\vartheta_{k-i} - f'(x_{k-i})), \tag{6.6}$$

$$i = 0, 1, \ldots, n-1.$$

Notons $B_k = D_k - f''(x_k)$. En procédant comme pour le lemme 3.1, on aboutit à l'estimation

$$\|B_k r_{k-i}\| \leqslant h_{k-i} \|r_{k-i}\| + \|\vartheta_{k-i} - f'(x_{k-i})\| + \|\varphi_{k-i} - f'(y_{k-i})\|,$$

où $h_{k-i} \to 0$ pour $k \to \infty$. Compte tenu de (6.4) et (6.5) il en résulte

$$\|B_k r_{k-i}\| \leqslant h_{k-i} \|r_{k-i}\| + C_5 \|r_{k-i}\|^t, \qquad C_5 < \infty,$$

ou

$$\| B_k r_{k-i} \| \leqslant \tilde{h}_{k-i} \| r_{k-i} \|,$$

où $\tilde{h}_{k-i} = h_{k-i} + C_5 \| r_{k-i} \|^{t-1} \to 0$ pour $k \to \infty$.

La démonstration ultérieure reproduit les raisonnements du lemme 3.1.

Définissons le vecteur

$$g_k = \left(\frac{f(x_k + \rho_k v_1) - f(x_k)}{\rho_k}, \ldots, \frac{f(x_k + \rho_k v_n) - f(x_k)}{\rho_k} \right), \tag{6.7}$$

où $|\rho_k| \leqslant |\mu_k|$ (si $\rho_k = \mu_k$, alors $g_k = \vartheta_k$).

Il est clair que la convergence et la vitesse de convergence de la suite (6.1) dépendent, en plus de la valeur de D_k, de la qualité de l'approximation de $f'(x_k)$ par g_k. Pour garantir une grande vitesse de convergence de (6.1) vers la solution on doit avoir (on le verra dans la suite de l'exposé), pour tout k, les inégalités

$$0 < |\rho_k| \leqslant \xi_k \| p_k \|, \tag{6.8}$$

avec ξ_k tendant de façon quelconque vers zéro pour $k \to \infty$. Si à une itération la valeur choisie de ρ_k ne satisfait pas aux conditions (6.8), on fractionne ρ_k, on calcule un nouveau vecteur g_k, puis un nouveau p_k et on teste (6.8). Comme $g_k \to f'_k$ pour $|\rho_k| \to 0$ et qu'alors $\| p_k \| \to \| D_k^{-1} f'_k \|$ et $\| D_k^{-1} f'_k \| > 0$ quel que soit $x_k \neq x_*$ (D_k^{-1} est régulière en tant qu'inverse de D_k ; en ce qui concerne le calcul de D_k^{-1} voir n° 3 (Calcul du vecteur p_k)), les conditions (6.8) ont lieu pour ρ_k suffisamment petits.

Détermination de la direction de déplacement. En se donnant une valeur γ_0 (on la choisit naturellement suffisamment petite), on calcule les valeurs de $f(x)$ aux points $x_k \pm \gamma_0 D_k^{-1} g_k$. Si en l'un de ces points la valeur de la fonction est inférieure à $f(x_k)$, on prend pour p_k le vecteur correspondant $D_k^{-1} g_k$ ou $-D_k^{-1} g_k$ (la condition (6.2) ayant lieu en vertu de la convexité de $f(x)$). Si les deux valeurs de la fonction sont plus grandes que $f(x_k)$, on fractionne γ_0 tant que l'une d'elles ne devient pas inférieure à $f(x_k)$; le vecteur correspondant est pris pour p_k.

Il se peut cependant que pour γ petits la fonction ne décroisse suivant aucune direction $\pm D_k^{-1} g_k$. Cela signifie que 1) les valeurs de γ pour lesquelles la fonction décroît ne sont pas encore atteintes ou 2) on a la condition $(f'_k, D_k^{-1} g_k) = 0$ (ce qui n'est possible, on le verra par la suite, qu'au stade initial du processus auquel cas aucun des vecteurs $\pm D_k^{-1} g_k$ n'est évidemment pris pour p_k). Pour exclure cette éventualité on calcule un nouveau vecteur $g_{k,1}$ en modifiant ρ_k (mais en conservant les conditions (6.8)), puis un nouveau $D_k^{-1} g_{k,1}$ et, à partir d'un certain $\gamma \leqslant \gamma_0$, on recherche également les valeurs de la fonction aux points $x_k \pm \gamma D_k^{-1} g_{k,1}$. Si $x_k \neq x_*$, l'une des direc-

tions $\pm D_k^{-1} g_k$ ou $\pm D_k^{-1} g_{k,1}$ est nécessairement une direction de descente, et le vecteur correspondant est pris pour p_k.

ALGORITHME DE CHOIX DU PAS. On choisira α_k comme suit : on pose

$$\bar{\alpha}_k = \min \left\{ 1,\ R \frac{|(g_k, p_k)|}{\| p_k \|^3} \right\}, \tag{6.9}$$

où $0 < R < \infty$, et on vérifie l'inégalité

$$f(x) - f(x_k) \leqslant \varepsilon \alpha^2 \beta_k (g_k, p_k), \tag{6.10}$$

avec $x = x_k + \alpha p_k$, $\beta_k = -\mathrm{sg}\,(g_k, p_k)$, $0 < \varepsilon < 1/2$.

Si cette inégalité a lieu pour $\alpha = \bar{\alpha}_k$, celle-ci est prise pour valeur cherchée. Dans le cas contraire, on fractionne $\bar{\alpha}_k$ jusqu'à ce que (6.10) soit vraie et on prend pour valeur cherchée la valeur de α_k ainsi obtenue.

Ce procédé de choix de α_k suppose évidemment $(g_k, p_k) \neq 0$. Si cette inégalité devient une égalité à une itération (cela ne peut arriver qu'au début du processus), on fractionne ρ_k et on recalcule g_k.

Analysons les propriétés de la suite (6.1) lorsqu'on construit comme indiqué la matrice D_k, le vecteur g_k et le paramètre α_k.

THÉORÈME 6.1. *Si $f(x)$ est une fonction deux fois continûment dérivable satisfaisant aux conditions* (2.4), *la matrice D_k se définit pour tout $k \geqslant n - 1$ par le système* (6.3), *le vecteur g_k par l'expression* (6.7), *avec ρ_k vérifiant les conditions* (6.8), *et α_k comme décrit plus haut, on a pour la suite* (6.1) *des assertions analogues à celles du théorème* 3.1.

DÉMONSTRATION. Pour utiliser le résultat du lemme 6.1, il faut montrer en premier lieu que dans les conditions du théorème la suite en question vérifie la condition $\| x_{k+1} - x_k \| \to 0$.

Développons au second ordre $f(x)$ autour de x_k, il vient

$$f_{k+1} - f_k = \alpha_k \beta_k (g_k, p_k) \left[\frac{(f'_k, p_k)}{\beta_k (g_k, p_k)} + \frac{\alpha_k}{2} \frac{(f''_{kc} p_k, p_k)}{\beta_k (g_k, p_k)} \right],$$

où $x_{kc} = x_k + \theta (x_{k+1} - x_k)$, $0 \leqslant \theta \leqslant 1$. Etant donné $\beta_k (g_k, p_k) < 0$, l'inégalité (6.10) a lieu si

$$\frac{(f'_k, p_k)}{\beta_k (g_k, p_k)} + \frac{\alpha_k}{2} \frac{(f''_{kc} p_k, p_k)}{\beta_k (g_k, p_k)} \geqslant \alpha_k \varepsilon$$

ou, ce qui revient au même,

$$\frac{1}{\alpha_k} \frac{(f'_k, p_k)}{\beta_k (g_k, p_k)} + \frac{1}{2} \frac{(f''_{kc} p_k, p_k)}{\beta_k (g_k, p_k)} \geqslant \varepsilon. \tag{6.11}$$

Par suite de (6.2) et du choix de β_k,

$$\frac{(f'_k, p_k)}{\beta_k (g_k, p_k)} > 0.$$

Par conséquent, on a, pour un certain $\alpha_k > 0$, l'inégalité (6.11) et donc (6.10), ce qui démontre la possibilité de choisir α_k comme décrit plus haut.

Ainsi, $f_{k+1} < f_k$ en vertu de (6.10). Cela signifie que $x_k \in S = \{x : f(x) \leqslant f(x_0)\}$ pour tout k, et, $f(x)$ étant minorée, $f_k - f_{k+1} \to 0$. Cela fait que (6.10) entraîne pour $k \to \infty$

$$\alpha_k^2 \,|\, (g_k, p_k) \,| \to 0. \tag{6.12}$$

Puisque $\alpha_k \leqslant \bar{\alpha}_k$, il découle de (6.9)

$$|\, (g_k, p_k) \,| \geqslant \frac{\alpha_k}{R} \,\| p_k \|^3.$$

Compte tenu de la dernière inégalité la condition (6.12) montre que $\| x_{k+1} - x_k \| = \alpha_k \| p_k \| \to 0$ pour k tendant vers l'infini. Les conditions du théorème garantissent donc les exigences du lemme 6.1 et donc

$$\| D_k - f_k'' \| \to 0. \tag{6.13}$$

Montrons que dans les conditions du théorème pour $k \to \infty$

$$\frac{(f_k', p_k)}{\beta_k (g_k, p_k)} \to 1. \tag{6.14}$$

On a

$$\frac{(f_k', p_k)}{\beta_k (g_k, p_k)} = \frac{1}{\beta_k} + \frac{(f_k' - g_k, p_k)}{\beta_k (g_k, p_k)} \leqslant \frac{1}{\beta_k} + \frac{\| f_k' - g_k \| \, \| p_k \|}{|\, (g_k, p_k) \,|}. \tag{6.15}$$

Le vecteur g_k s'évalue par

$$\| g_k - f_k' \| \leqslant C_6 \,|\, \rho_k \,|\, n^{1/2} = C_7 \,|\, \rho_k \,| \tag{6.16}$$

(une estimation analogue à (6.4)). Puisque

$$|\, (g_k, p_k) \,| = |\, (D_k p_k, p_k) \,|, \tag{6.17}$$

il en découle, par suite des conditions (6.13) et (2.4), qu'à partir d'une itération

$$|\, (g_k, p_k) \,| \geqslant m_1 \,\| p_k \|^2, \tag{6.18}$$

où $0 < m_1 \leqslant m$. On s'assure moyennant les estimations (6.8), (6.16) et (6.18) qu'à partir d'une itération on a les conditions

$$\frac{\| f_k' - g_k \| \, \| p_k \|}{|\, (g_k, p_k) \,|} \leqslant \frac{C_7 \,|\, \rho_k \,| \, \| p_k \|}{m_1 \,\| p_k \|^2} \leqslant \frac{C_7}{m_1} \xi_k \to 0.$$

Il résulte par suite de (6.15) $\beta_k = +1$ à partir d'une certaine itération (vu la positivité du second membre de (6.15)) et donc, en effet, la condition (6.14).

Le gradient $f'(x)$ est borné sur l'ensemble S: $\| f_k' \| \leqslant L$. Vu que $|\, \rho_k \,| \leqslant \rho < \infty$ on se convainc à l'aide de (6.16) que $\| g_k \| \leqslant L_1$ pour tout k. Par analogie avec le théorème 3.1 on établit que $\| D_k^{-1} \| \leqslant$

$\leqslant M_2$ quel que soit $k \geqslant n-1$. Par conséquent,

$$\| p_k \| \leqslant \|D_k^{-1}\| \, \| g_k \| \leqslant C_8.$$

Cette estimation et l'inégalité (6.18) aidant, on établit que pour k suffisamment élevés

$$\frac{|(g_k, p_k)|}{\| p_k \|^3} \geqslant \frac{m_1 \| p_k \|^2}{\| p_k \|^3} \geqslant \frac{m_1}{\| p_k \|} \geqslant C_9 > 0. \tag{6.19}$$

Par suite, (6.9) entraîne, à partir d'un certain k,

$$\bar{\alpha}_k \geqslant \bar{\alpha} > 0. \tag{6.20}$$

En utilisant les conditions (6.14) et (6.18) on constate de même sans peine que pour k suffisamment importants l'inégalité (6.11) et, partant, (6.10) sont remplies pour les valeurs $\alpha \geqslant \tilde{\alpha} > 0$. Réunies à l'estimation (6.20) ces inégalités montrent que

$$\alpha_k \geqslant C_{10} > 0$$

à partir d'un certain k. Cette estimation fait que la condition $\alpha_k \| p_k \| \to 0$ dont nous avons parlé plus haut implique, quand $k \to \infty$,

$$\| p_k \| \to 0. \tag{6.21}$$

Du moment que $\| g_k \| = \| D_k p_k \| \leqslant M_1 \| p_k \|$, on a pour $\| p_k \|$ tendant vers zéro

$$\| g_k \| \to 0. \tag{6.22}$$

En vertu des conditions (6.8), (6.16), (6.21) et (6.22) on dit que, lorsque $k \to \infty$,

$$\| f'(x_k) \| \to 0,$$

ce qui signifie (conséquence de l'inégalité (1.12) juste pour des fonctions fortement convexes) la convergence vers la solution de la suite (6.1).

Evaluons la vitesse de convergence de la méthode.

Par suite des conditions (6.17), (6.21) et de la continuité uniforme des dérivées secondes de la fonction sur l'ensemble S, on a pour $k \to \infty$

$$\frac{(\tilde{f}_{kc} p_k, p_k)}{|(g_k, p_k)|} \to 1.$$

Utilisons cette condition et (6.14). On s'assure aisément que l'inégalité (6.11) et donc (6.10) sont vraies pour $\alpha = 1$ si k est suffisamment grand. Les relations (6.19) entraînent pour $\| p_k \| \to 0$

$$\frac{|(g_k, p_k)|}{\| p_k \|^3} \to \infty.$$

C'est pourquoi, $\bar{\alpha}_k = 1$ à partir d'un certain k lorsqu'on choisit $\bar{\alpha}_k$ par la condition (6.9).

A la lumière de ces remarques, $\alpha_k \equiv 1$ à partir d'une certaine itération et, de plus,

$$x_{k+1} - x_k = -D_k^{-1} g_k.$$

D'autre part, on peut trouver une matrice $\widetilde{D}_k^{-1}$ telle que

$$x_{k+1} - x_k = -\widetilde{D}_k^{-1} f'_k.$$

Dans les conditions (6.8) et (6.16) la suite de matrices $\widetilde{D}_k$ peut être choisie de façon qu'on ait

$$\widetilde{D}_k \to D_k. \tag{6.23}$$

A cet effet, on pose par exemple

$$\widetilde{D}_k = D_k - \frac{(f'_k - g_k)}{\| x_{k+1} - x_k \|^2} (x_{k+1} - x_k)^*.$$

Il n'est plus difficile de démontrer la convergence superlinéaire de la suite (6.1). Pour y arriver, on raisonne comme pour le théorème 2.1 et on démontre l'inégalité

$$\| x_{k+1} - x_* \| \leqslant \| \widetilde{D}_k^{-1} \| \, \| \widetilde{D}_k - f''_{kc} \| \, \| x_k - x_* \|.$$

Utilisons ensuite les conditions (6.13), (6.23) et la continuité des dérivées secondes afin de nous convaincre que, quand $k \to \infty$, $\| \widetilde{D}_k - f''_{kc} \| \to 0$ et $\| \widetilde{D}_k^{-1} \|$ est une quantité bornée. Par suite, on a pour k tendant vers l'infini

$$\| x_{k+1} - x_* \| \leqslant \lambda_k \| x_k - x_* \|, \tag{6.24}$$

où $\lambda_k \to 0$, ce qui démontre la convergence superlinéaire de $\{x_k\}$ et du même coup le théorème.

3. Quelques remarques sur la mise en œuvre des méthodes des directions duales. DIVERS ALGORITHMES. Les vecteurs r_k servant à construire la matrice D_k sont assujettis aux mêmes conditions que dans le cas de la suite (3.5). Tout ce que nous avons dit au n° 3 § 3 sur la construction de divers algorithmes du type (3.4) reste donc vrai pour le processus (6.1).

CALCUL DU VECTEUR p_k. Les résultats du n° 4 § 3 s'appliquent en l'occurrence entièrement. Ainsi, la base $\bar{s}_{k+1}, \bar{s}_k, \ldots, \bar{s}_{k-n+2}$ duale de $\psi_{k+1}, \psi_k, \ldots, \psi_{k-n+2}$ se forme selon les formules (analogues à (3.21))

$$\bar{s}_{k+1} = \frac{s_{k-n+1}}{(s_{k-n+1}, e_{k+1})}, \quad \bar{s}_{k+1-j} = s_{k+1-j} -$$
$$- (s_{k+1-j}, e_{k+1}) \bar{s}_{k+1}, \quad j = 1, \ldots, n-1.$$

Pour vérifier l'indépendance linéaire des vecteurs $\psi_{k+1}, \psi_k, \ldots, \psi_{k-n+2}$ il suffit de calculer le produit scalaire (s_{k-n+1}, ψ_{k+1}): $(s_{k-n+1}, \psi_{k+1}) \neq 0$

confirme cette propriété. Dans le cas contraire, il faut modifier ou le vecteur r_{k+1} ou l'un quelconque des vecteurs ϑ_{k+1}, φ_{k+1} (cela fait varier ψ_{k+1}).

Dans la pratique, les approximations successives sont à générer par la formule (analogue à (3.25))

$$x_{k+1} = x_k - \alpha_k \sum_{i=0}^{n-1} (s_{k-i}, \quad g_k)\, r_{k-i}. \tag{6.25}$$

Démarrage du processus. Les premières itérations du processus ($k < n-1$) s'effectuent de différentes manières. On peut par exemple suivre l'une des directions $\beta_k g_k$, $\beta_k = \pm 1$ en choisissant le signe de β_k de façon que $f(x)$ décroisse.

Afin d'assurer l'uniformité du processus itératif (6.25), on peut le faire débuter comme l'indique le n° 5 § 3.

Minimisation d'une forme quadratique. Soit $f(x) = \frac{1}{2}\ (Ax,x) + (b,x) + c$, avec $(Ax,x) > 0$ pour tout $x \neq 0$. On constate alors aisément que le vecteur $\vartheta_k = g_k = f'(x_k)$, $\varphi_k = f'(y_k)$, $\psi_k = e_k$, i.e. $D_k = A_k$, et le processus (6.1) coïncide avec (3.4). Par conséquent (voir n° 6 § 3), le premier processus permet de trouver le minimum d'une fonction quadratique en n pas, ce qui demande le calcul de $(n+1)^2$ valeurs de la fonction.

Sur le choix du vecteur g_k. La méthode (6.1) suppose, en plus de l'approximation de la matrice $f''(x)$, le remplacement du gradient $f'(x)$ par son analogue aux différences, à savoir par le vecteur g_k. La convergence superlinéaire a alors lieu (comme nous l'avons noté) dans les conditions (6.8). Si on n'a celles-ci qu'à condition de calculer plusieurs fois g_k à une itération, le coût du processus (surtout pour un espace de dimension élevée) augmente en conséquence.

Notons la possibilité de choisir $|\rho_k| = \|p_{k-1}\|^2$ si $\|p_k\| < \|p_{k-1}\|$ à chaque itération. Il se peut très bien que pour un tel choix de ρ_k, la dernière inégalité (6.8) est vérifiée au moins à partir d'une certaine itération. En effet, la vitesse de convergence s'évalue finalement par (6.24). La convergence ainsi évaluée est d'habitude plus lente que la convergence quadratique

$$\| x_{k+1} - x_k \| \leqslant \| x_k - x_{k-1} \|^2, \quad \| x_k - x_{k-1} \| \to 0,$$

c'est-à-dire, étant donné l'estimation (6.24), on a en général $\| p_{k-1} \|^2 < \| p_k \|$ (rappelons que $\alpha_k = 1$, i.e. $p_k = x_{k+1} - x_k$, à partir d'un certain k). En choisissant la suite $\{\zeta_k\}$ de façon que $\zeta_k \to 0$ assez lentement, on peut donc espérer qu'avec le choix $\rho_k = \| p_{k-1} \|^2$ l'inégalité (6.8) a lieu sans qu'il soit nécessaire de calculer plusieurs fois g_k. Si la condition (6.8) n'est pas remplie de suite (i.e. on a à frac-

tionner ρ_k), cela atteste une convergence voisine de celle quadratique.

Nous terminons ce numéro en disant que moyennant les résultats du n° 3 § 3 et du présent paragraphe on établit les conditions de convergence de la méthode de Newton modifiée n'exigeant pas le calcul de dérivées.

4. Méthodes des directions conjuguées. Voyons un procédé de construction de directions conjuguées dont l'idée est toute différente de celle des méthodes considérées au § 4.

Soit encore

$$f(x) = \frac{1}{2}(Ax, x) + (b, x) + c,$$

où $(Ax, x) > 0$ pour tout $x \neq 0$. Admettons que les directions (non nulles) $p_1, \ldots, p_m$, $m < n$, sont A-conjuguées et soient $E^m(x_0)$ et $E^m(x_{0,m})$ deux sous-espaces m-dimensionnels distincts de E^n engendrés par les vecteurs $p_1, \ldots, p_m$ et passant par les points x_0 et $x_{0,m}$. Si x_m et $x_{m,m}$ réalisent alors le minimum de $f(x)$ sur ces sous-espaces, on a

$$(f'(x_m), p_i) = 0,$$
$$(f'(x_{m\ m}), p_i) = 0, \ i = 1, 2, \ldots, m.$$

Par conséquent, $(f'(x_m) - f'(x_{m,m}), p_i) = 0$ ou

$$(A(x_m - x_{m,m}), p_i) = 0, \ i = 1, 2, \ldots, m.$$

Ainsi, les points de minimum de $f(x)$ étant définis sur des sous-espaces distincts engendrés par les directions $p_1, \ldots, p_m$ A-conjuguées, la direction $p_{m+1} = x_{m,m} - x_m$ se trouve conjuguée des $p_1, \ldots, p_m$.

Le procédé décrit de construction de vecteurs conjugués n'exige pas qu'on calcule le gradient ou qu'on l'approche par son analogue aux différences. Donnons maintenant un algorithme concret de minimisation d'une fonction quadratique, qui fait appel à cette technique pour construire les vecteurs conjugués.

Choisissons un point x_0 quelconque et un vecteur p_1 (qui est quelconque lui aussi). La m-ième itération de l'algorithme ($m = 1, 2, \ldots, n$) comprend les phases suivantes:

1) on calcule le point

$$x_m = x_{m-1} + \alpha_m p_m, \tag{6.26}$$

avec α_m défini à partir de la condition de minimum de la fonction

$$f(\alpha) = f(x_{m-1} + \alpha p_m);$$

2) on calcule le point

$$x_{0,m} = x_m + r_m, \tag{6.27}$$

r_m étant un vecteur arbitraire qui ne constitue pas une combinaison linéaire des vecteurs $p_1, \ldots, p_m$ (plus loin le choix de r_m sera examiné plus en détail);

3) on calcule les points

$$x_{k,m} = x_{k-1,m} + \alpha_{k,m} p_k, \quad k = 1, \ldots, m,$$

$\alpha_{k,m}$ étant défini par la condition de minimum de $f(\alpha) = f(x_{k-1,m} + \alpha p_k)$;

4) on définit le vecteur $p_{m+1} = x_{m,m} - x_m$. Sur ce, la m-ième itération prend fin.

Le vecteur r_m (dans (6.27)) ne doit pas faire partie du sous-espace $E^m(x_0)$, cela pour que $x_{0,m} \not\subset E^m(x_0)$. Puisque x_m réalise le minimum de $f(x)$ sur ce sous-espace, il est clair qu'aucun vecteur $x - x_m$ qui est une direction de décroissance de $f(x)$ n'appartient à $E^m(x_0)$. On peut donc prendre comme r_m toute direction de descente de $f(x)$ issue du point x_m. Il est en particulier commode de choisir r_m coïncidant avec l'un des axes de coordonnées; si r_m ainsi choisi n'est pas une direction de descente, on le fait confondre avec l'autre axe.

Conformément aux résultats du § 4, le point x_n obtenu par la formule (6.26) est un minimum de $f(x)$: $x_n = x_*$. On trouve x_* en résolvant $1 + 2 + \ldots + n = \frac{n+1}{2} n$ fois les problèmes de minimisation à une variable (pour définir les facteurs α_m et $\alpha_{k,m}$).

Avec cette approche de la construction de directions conjuguées, on forme divers algorithmes de minimisation de fonctions non quadratiques. Il va sans dire que dans tout algorithme de ce type les directions $p_1, \ldots, p_m$, $m \leqslant n$, ne sont plus conjuguées (voir à ce propos n° 1 § 5). On s'attend cependant à ce que de bonnes méthodes permettent de construire, dans un voisinage suffisamment petit du minimum x_* (d'une fonction régulière convexe), des vecteurs possédant des propriétés proches de celles des vecteurs conjugués. De tels algorithmes peuvent se révéler efficaces dans la recherche du minimum de fonctions non quadratiques.

L'algorithme ci-dessus se base justement sur les considérations énoncées.

Soient $x_{1,0}$ un point quelconque et $v_{1,1}, \ldots, v_{1,n}$ un repère orthonormé; la k-ième itération de l'algorithme ($k = 1, 2, \ldots$) consiste à

1) calculer pour $i = 1, 2, \ldots, n$

$$x_{k,i} = x_{k,i-1} + \alpha_{k,i} v_{k,i},$$

où $\alpha_{k,i}$ se définissent par la condition de minimum de la fonction

$$f(\alpha) = f(x_{k,i-1} + \alpha v_{k,i});$$

2) poser $v_{k,n+1}=\frac{x_{k,n}-x_{k,0}}{\gamma_k}$, avec $\gamma_k=\|x_{k,n}-x_{k,0}\|$, et trouver le point $x_{k,n+1}=x_{k,n}+\alpha_{k,n+1}v_{k,n+1}$, où $\alpha_{k,n+1}$ est défini par la condition de minimum de la fonction

$$f(x_{k,n}+\alpha v_{k,n+1});$$

3) supposer $\alpha_{k,s}=\max\{\alpha_{k,i}: i=1, 2, \ldots, n\}$, Δ_k un déterminant dont les colonnes sont les vecteurs $v_{k,1}, \ldots, v_{k,n}$ et $\varepsilon>0$ une constante positive petite. Si

$$\frac{\alpha_{k,s}\Delta_k}{\gamma_k}\geqslant\varepsilon,$$

on pose $v_{k+1,i}=v_{k,i}$ pour $i\neq s$ et $v_{k+1,s}=v_{k,n+1}$, auquel cas on a

$$\Delta_{k+1}=\frac{\alpha_{k,s}\Delta_k}{\gamma_k}. \tag{6.28}$$

Mais si $\frac{\alpha_{k,s}\Delta_k}{\gamma_k}<\varepsilon$, on pose $v_{k+1,i}=v_{k,i}$ pour tous les $i=1, 2, \ldots, n$; ceci étant, $\Delta_{k+1}=\Delta_k$;

4) poser $x_{k+1,0}=x_{k,n+1}$.

C'est la dernière étape de la k-ième itération.

L'égalité (6.28) est à prouver. Avant de le faire, discutons l'algorithme proposé.

Considérons une version simplifiée dont la k-ième itération s'effectue comme suit.

1) On construit les points $x_{k,i}$, $i=1, 2, \ldots, n$, comme dans 1) de l'algorithme originel.

2) On calcule $x_{k,n+1}=x_{k,n}+\alpha_{k,n+1}v_{k,n+1}$, où $v_{k,n+1}=x_{k,n}-x_{k,0}$ et $\alpha_{k,n+1}$ minimise $f(x_{k,n}+\alpha v_{k,n+1})$.

3) On pose $v_{k+1,i}=v_{k,i+1}$, $i=1, 2, \ldots, n$.

4) On pose $x_{k+1,0}=x_{k,n+1}$.

Soit $k=2$. On a alors

$$x_{2,0}=x_{1,n+1}=x_{1,n}+\alpha_{1,n+1}v_{1,n+1},$$

$$x_{2,n}=x_{2,n-1}+\alpha_{2,n}v_{2,n}=x_{2,n-1}+\alpha_{2,n}v_{1,n+1},$$

i.e. $x_{2,0}$ et $x_{2,n}$ sont deux points de minimum de $f(x)$ sur un espace unidimensionnel (engendré par le vecteur $v_{1,n+1}$) passant par deux points différents $x_{1,n}$ et $x_{2,n-1}$. Si $f(x)$ est une fonction quadratique, la direction $v_{2,n+1}=x_{2,n}-x_{2,0}$ se trouve conjuguée de la direction $v_{1,n+1}=v_{2,n}$ comme cela découle des résultats antérieurs. En continuant dans cette voie, on se convainc que si les vecteurs $v_{k,1}, \ldots, v_{k,n}$ sont linéairement indépendants pour tout $k=1,2, \ldots, n$, alors $v_{k,n+1}, v_{k,n}, \ldots, v_{k,n-k+2}$ s'avèrent conjugués après la k-ième itération, i.e. on obtient n vecteurs conjugués en n itérations. Avec ce procédé de formation des $v_{k,1}, \ldots, v_{k,n}$ il est impossible de garantir leur indépendance linéaire. En effet, si on a $\alpha_{k,1}=0$

pour un k, alors

$$v_{k,n+1} = x_{k,n} - x_{k,0} = x_{k,n} - x_{k,1} = \sum_{i=2}^{n} \alpha_{k,i} v_{k,i}$$

(on le constate facilement), i.e. à la $(k+1)$-ième itération le système de vecteurs $v_{k+1,i} = v_{k,i+1}$, $i = 1, 2, \ldots, n$, possède la propriété d'être linéairement dépendant. Il est impossible dans ce cas de construire un système de n vecteurs conjugués ; cela veut dire qu'avec l'algorithme simplifié on ne garantit pas la résolution même pour une fonction quadratique. Si les étapes 2) et 3) de l'algorithme principal sont compliquées, c'est justement dans le but d'éviter que les vecteurs $v_{k,i}$, $i = 1, 2, \ldots, n$, soient linéairement dépendants (il se trouve que $\Delta_k \geqslant \varepsilon$).

Il faut cependant noter qu'avec l'algorithme originel on ne garantit plus la résolution en n itérations du problème de la recherche du minimum d'une fonction quadratique. En effet, en passant du système $v_{k,1}, \ldots, v_{k,n}$ au système $v_{k+1,1}, \ldots, v_{k+1,n}$ on risque de remplacer (voir 3)) l'un des vecteurs conjugués construits si bien qu'on ne garantit plus l'obtention de n vecteurs conjugués en un nombre fini d'itérations. Le système de vecteurs $v_{k,i}$ peut de plus ne pas changer en passant à la $(k+1)$-ième itération.

Prouvons l'égalité (6.28) :

$$\Delta_{k+1} = \det[(v_{k+1,1}, \ldots, v_{k+1,n})] =$$
$$= \det[(v_{k,1}, \ldots, v_{k,s-1}, v_{k,n+1}, v_{k,s+1}, \ldots, v_{k,n})].$$

Mais

$$v_{k,n+1} = \frac{1}{\gamma_k}(x_{k,n} - x_{k,0}) = \frac{1}{\gamma_k} \sum_{i=1}^{n} \alpha_{k,i} v_{k,i}.$$

Donc

$$\Delta_{k+1} = \frac{\alpha_{k,s}}{\gamma_k} \det[(v_{k,1}, \ldots, v_{k,s}, \ldots, v_{k,n})] = \frac{\alpha_{k,s}\Delta_k}{\gamma_k}.$$

Ainsi, on a $\Delta_k \geqslant \varepsilon$ quel que soit k, ce qui garantit justement l'indépendance linéaire des $v_{k,1}, \ldots, v_{k,n}$.

Etudions certaines propriétés de l'algorithme en question.

THÉORÈME 6.2. *Etant donné une fonction $f(x)$ strictement convexe continûment dérivable telle que l'ensemble $S = \{x : f(x) \leqslant f(x_{1,0})\}$ soit borné pour un choix arbitraire du point $x_{1,0}$, la suite*

$$\{x_{k,i}\},\ i = 0, 1, \ldots, n,\ k = 1, 2, \ldots, \tag{6.29}$$

construite de façon décrite converge vers le point de minimum de $f(x)$.

DÉMONSTRATION. L'existence et l'unicité du minimum x_* de $f(x)$ découlent sous ces hypothèses des résultats des lemmes I.3.1 et

I.3.4. La seule chose à démontrer est donc la convergence de la suite $\{x_{k,i}\}$. Tout point de (6.29) appartient à S ($x_{k,i} \in S$) parce que $f(x_{k,i}) = \min_{\alpha} f(x_{k,i-1} + \alpha v_{k,i}) \leqslant f(x_{k,i-1})$ et $f(x_{k+1,0}) = f(x_{k,n+1}) \leqslant f(x_{k,n})$. L'ensemble S est borné, i.e. il est compact (dans E^n). On peut donc extraire de toute suite infinie d'éléments $\in S$ une sous-suite convergente (vers un élément de S). En considérant $\{x_{k,i}\}$ pour $i = 0, 1, \ldots, n$ fixe, il existe, conformément à ce qu'on a dit, une suite infinie $\{x_{k_m,i}\}$ dont la limite est un point $x_i \in S$. La fonction $f(x)$ étant minorée, on a dans ce cas $f(x_{k_m,i+1}) \to f(x_{k_m,i})$, d'où, compte tenu de la continuité de $f(x)$, les égalités

$$f(x_{i+1}) = \lim_{k_m \to \infty} f(x_{k_m,i+1}) = \lim_{k_m \to \infty} f(x_{k_m,i}) = f(x_i). \qquad (6.30)$$

Montrons que $x_{i+1} = x_i$ pour $i = 0, 1, \ldots, n-1$. Par construction $\| v_{k,i} \| = 1$ quels que soient k et i; on assimile donc les vecteurs $v_{k,i}$ aux éléments de la sphère unité (un ensemble borné), et il existe par suite pour tout $i = 1, \ldots, n$ fixe une sous-suite $\{v_{k_m,i}\}$ qui converge vers un vecteur v_i. Puisque $x_{k,i+1} = x_{k,i} + \alpha_{k,i+1} v_{k,i+1}$ et $x_{k_m,i+1} \to x_{i+1}$, $x_{k_m,i} \to x_i$, $v_{k_m,i+1} \to v_{i+1}$, on a

$$x_{i+1} = x_i + \alpha_{i+1} v_{i+1}, \qquad i = 0, 1, \ldots, n-1,$$

où $\alpha_{i+1} = \lim_{k_m \to \infty} \alpha_{k_m,i+1}$. Vu que le point $x_{k,i+1}$ vérifie la condition $f(x_{k,i+1}) = \min_{\alpha} f(x_{k,i} + \alpha v_{k,i+1})$, on doit avoir

$$f(x_{i+1}) = \min_{\alpha} f(x_i + \alpha v_{i+1}), \quad i = 0, 1, \ldots, n-1, \qquad (6.31)$$

i.e. $f(x)$ atteint son minimum dans la direction v_{i+1} au point x_{i+1}. Mais il résulte de (6.30) que $f(x_{i+1}) = f(x_i)$. Etant donné la convexité stricte de $f(x)$, il existe un seul minimum dans la direction v_{i+1}, d'où $x_{i+1} = x_i$.

Ainsi, $x_0 = x_1 = \ldots = x_n$. Désignons par $\tilde{x}$ ce point commun. Les conditions (6.31) s'écrivent alors

$$f(\tilde{x}) \leqslant f(\tilde{x} + \alpha v_i), \; i = 1, 2, \ldots, n, \qquad (6.32)$$

α étant quelconque. S'agissant d'une fonction dérivable, ces conditions équivalent à

$$(f'(\tilde{x}), v_i) = 0, \; i = 1, 2, \ldots, n. \qquad (6.33)$$

Notons que $\det[(v_1, \ldots, v_n)] \geqslant \varepsilon$ du moment que $\det[(v_{k,1}, \ldots, v_{k,n})] \geqslant \varepsilon$, d'où l'indépendance linéaire des $v_1, \ldots, v_n$. Ce fait établi, (6.33) entraîne $f'(\tilde{x}) = 0$. $f(\tilde{x})$ étant strictement convexe, cela signifie que $\tilde{x}$ réalise le minimum de $f(x)$: $\tilde{x} = x_*$.

Ainsi, nous avons démontré l'existence d'une sous-suite $\{x_{k_m,\,i}\}$ qui converge vers le point x_*. Mais comme $f(x_{k+1,\,i}) \leqslant f(x_{k,\,i})$ pour tout $i = 0, 1, \ldots, n$ fixe et $f(x)$ est minorée, on a les conditions

$$\lim_{k\to\infty} f(x_{k,\,i}) = \lim_{k_m\to\infty} f(x_{k_m,\,i}) = f(x_i) = f(\tilde{x}).$$

Il en résulte, pour i fixe, les propriétés minimisantes de la suite $\{x_{k,\,i}\}$ et, partant, celles de la suite (6.29) qui converge donc, le minimum étant unique, vers le point x_*, c.q.f.d.

En démontrant les conditions (6.32) nous n'avons pas utilisé la dérivabilité de la fonction $f(x)$, i.e. les inégalités en question restent valables pour une fonction continue strictement convexe. Dans ce cas $\tilde{x}$, point limite de la suite (6.29), peut ne pas réaliser le minimum de $f(x)$ ((6.29) peut en l'occurrence présenter plus d'un point limite).

5. Discussion des résultats. Commençons par dire que le domaine d'application de la méthode des directions conjuguées est plus vaste que celui des techniques utilisant des directions duales, ce qu'on constate aisément en comparant les contraintes imposées à la fonction à minimiser dans les théorèmes 6.1 et 6.2.

L'étude des propriétés de la méthode considérée des directions conjuguées est loin d'être terminée. C'est le cas, par exemple, de la vitesse de convergence de l'algorithme, dont on sait qu'elle doit être plus mauvaise (lorsqu'on minimise les fonctions d'une même classe) que celle des méthodes du § 5. En effet, il suffit de dire que l'algorithme étudié ne garantit pas l'obtention du minimum d'une fonction quadratique en n itérations (et en général en un nombre fini de pas), i.e. il ne garantit pas la construction d'un système de n vecteurs conjugués en un nombre fini d'itérations. Sous l'aspect vitesse de convergence les méthodes des directions duales à convergence superlinéaire sont donc plus bonnes.

Tâchons de comparer le nombre d'opérations par itération des algorithmes considérés.

Dans les méthodes du type (6.1) on effectue par itération, selon la variante (voir n° 3), $n + 1$ ou $2(n + 1)$ calculs de la fonction (pour définir la matrice D_k^{-1}) et $n + 1$ calculs de la fonction (pour construire le vecteur g_k); il se peut qu'à certaines itérations ou bien g_k se définisse sans qu'on ait besoin de connaître de nouvelles valeurs de la fonction (si $\rho_k = \mu_k$) ou bien le volume de calcul augmente de plusieurs fois selon la qualité de l'approximation du gradient par g_k. Le choix de la direction de déplacement et de la grandeur du pas demande lui aussi plusieurs calculs de la fonction.

Chaque itération de la méthode des directions conjuguées coûte $n + 1$ calculs du minimum de la fonction dans la direction de dépla-

cement. A supposer que la résolution d'un problème unidimensionnel de minimisation demande en moyenne 3 à 4 calculs de la fonction, le nombre de calculs de la fonction par itération est sensiblement le même dans les méthodes des deux types. Il est vrai qu'on ignore la précision avec laquelle on doit calculer le minimum suivant la direction de déplacement dans la méthode des directions conjuguées pour que les propriétés du processus se conservent. Du point de vue de l'influence sur la convergence on doit préférer l'algorithme de choix de α_k dans le processus (6.1).

Sous la condition d'être praticables, les méthodes du type (6.1) sont dans l'ensemble plus efficaces que la méthode des directions conjuguées, mais nous voulons insister une fois de plus sur son domaine d'application plus vaste.

En étudiant le processus (6.1) nous avons en fait établi que les erreurs de calcul en $O(\| r_k \|^t)$ sur le vecteur e_k (voir (6.4), (6.5), (6.6)) et les erreurs de calcul en $O(\zeta_k \| p_k \|)$ sur le vecteur $f'(x_k)$ (voir (6.16)) n'altèrent pas les propriétés (convergence, estimation de la vitesse de convergence) du processus (3.4). Dans la variante de (3.4) où $r_{k+1} = x_{k+1} - x_k$ les erreurs s'évaluent par des expressions différentes. On a donc dans (3.4) à partir d'un pas $\alpha_k = 1$

$$\| r_k \| = \| p_{k-1} \| = \| x_k - x_{k-1} \| = \| A_{k-1}^{-1} f'_{k-1} \| \geqslant m_1 \| f'_{k-1} \|.$$

Compte tenu de (1.12) on a $\| r_k \| \geqslant m_1 m \| x_{k-1} - x_* \|$.

Si $r_{k+1} = x_{k+1} - x_k$, les erreurs en $O(\| x_{k-1} - x_* \|^t)$ et en $O(\zeta_k \| x_k - x_* \|)$ résultant du calcul des vecteurs e_k et f'_k n'influent donc pas sur les propriétés du processus (3.4).

COMMENTAIRE

§ 1. L'idée de la méthode du gradient a été émise par *A. L. Cauchy* [49]. Divers auteurs dont *L. Kantorovitch* [19; 1], *H. Kelley* [80], *B. Poliak* [28; 1], *M. Altman* [38; 1], *Y. Liubitch* [23], *Y. Liubitch et G. Maïstrovski* [24] ont étudié les méthodes du type gradient dans l'optique différente. Voir la bibliographie (très fournie) des ouvrages cités.

La variante avec le pas choisi à partir de la condition (1.2) est proposée pour la première fois.

§ 2. La méthode de Newton pour les problèmes de minimisation et les équations figure dans *L. Kantorovitch* [19; 2], *L. Kantorovitch et G. Akilov* [78], *L. Collatz* [51]... *M. Yakovlev* [37] a démontré la convergence superlinéaire de la méthode généralisée si la grandeur du pas est choisie par la condition de minimum dans la direction de déplacement. *A. A. Goldstein et J. F. Price* [68], *J. W. Daniel* [52; 1], *Y. Daniline* [8; 1,2] ont étudié la méthode de Newton à pas variable avec le choix de α_k indépendant de la recherche du minimum dans la direction de déplacement.

§ 3. Ce paragraphe est rédigé d'après *Y. Daniline et B. Pchénitchny* [8; 1].

§ 4. La première méthode des directions conjuguées, celle du gradient conjugué, a été proposée par *M. R. Hestenes* et *E. Stiefel* [73] pour des problèmes d'algèbre linéaire. On doit à *W. C. Davidon* [54; 1] une autre approche de la construction des méthodes des directions conjuguées qui s'appliquent cette fois

à la minimisation d'une fonction quadratique. L'idée de cet auteur a été développée par *R. Fletcher et M. J. D. Powell* [61] et par d'autres savants.

Dans l'algorithme de *B. Pchénitchny* [29; 3] la construction de directions conjuguées n'est pas liée à la recherche du minimum de la fonction dans la direction de déplacement.

Pour certaines propriétés des directions conjuguées consulter *D. Faddeev et V. Faddeeva* [57]. Le procédé général de construction des méthodes des directions conjuguées a été élaboré par *H. Y. Huang* [75]. Parmi les résultats inédits citons la formule (4.45) et la méthode (4.63).

§ 5. *R. Fletcher et C. M. Reeves* [62] ont suggéré d'utiliser la méthode du gradient conjugué pour minimiser des fonctions non quadratiques. Pour la convergence de cette méthode et l'évaluation de sa vitesse voir *J. W. Daniel* [52; 1, 2], *B. Poliak* [28; 2], *G. Maïstrovski* [25; 1] [2], *S. Smoliak* [31]. La convergence de la méthode (4.48) et les estimations de la vitesse de convergence ont été établies par *M. J. D. Powell* [94; 3] (ces résultats sont cités dans *J. W. Daniel* [52; 1], *E. Polak* [93; 2]. La démonstration de la convergence des techniques des directions conjuguées suit *Y. Daniline* [7; 1].

§ 6. Les méthodes des directions duales sans calcul de dérivées ont été examinées dans *Y. Daniline et B. Pchénitchny* [8; 2]. Les méthodes des directions conjuguées ont intéressé entre autres *C. S. Smith* [97], *M. J. D. Powell* [94; 2], *W. I. Zangwill* [102; 1], *J. W. Daniel* [52; 1]. Les auteurs y ont puisé pour écrire le présent paragraphe. Un tour d'horizon des méthodes de minimisation sans calcul de dérivées a été fait par *R. P. Brent* [44].

CHAPITRE III

MÉTHODES DE RÉSOLUTION DE PROBLÈMES AVEC CONTRAINTES

Le présent chapitre expose diverses techniques de minimisation de fonctions en présence de contraintes imposées aux variables. On commence par des méthodes de résolution du problème de programmation quadratique qui constitue un problème auxiliaire dans divers algorithmes, puis on expose des algorithmes de résolution des problèmes de programmation convexe et non convexe. Partout où cela est possible on évalue la vitesse de convergence.

§ 1. Problème de programmation quadratique

On entend en général par problème de programmation quadratique la recherche du minimum d'une fonction quadratique avec contraintes linéaires. Ainsi, le problème de programmation quadratique consiste à minimiser

$$f(x) = \frac{1}{2}(x, Cx) + (d, x) \tag{1.1}$$

sous les contraintes

$$\begin{aligned} (a_i, x) - b_i &\leqslant 0, \quad i \in \mathcal{J}^-, \\ (a_i, x) - b_i &= 0, \quad i \in \mathcal{J}^0. \end{aligned} \tag{1.2}$$

Ici $x \in E^n$, $a_i \in E^n$, $i \in \mathcal{J}^- \cup \mathcal{J}^0$, $d \in E^n$, b_i sont des nombres, la $n \times n$-matrice C est symétrique définie positive, i.e. $(x, Cx) \geqslant 0$ pour tous les x et $\mathcal{J}^-$ et $\mathcal{J}^0$ sont deux ensembles finis d'indices.

La résolution numérique de ce problème se base sur la méthode du gradient conjugué dont l'idée principale est en l'occurrence la suivante.

Soit x_0 un point vérifiant les contraintes (1.2). Dégageons de (1.2) les contraintes saturées. Elles définissent une face d'un ensemble polyédrique décrit par les inégalités linéaires (1.2). Trouvons le minimum de $f(x)$ sur cette face par la méthode du gradient conjugué. Ou bien le point obtenu est solution du problème proposé ou bien on peut passer à une autre face et la procédure reprend par le com-

mencement. Vu que la méthode du gradient conjugué minimise $f(x)$ en un nombre fini de pas et que les faces sont en nombre fini, un tel algorithme converge au bout d'un nombre fini de pas.

1. Opérateurs de projection. Soit maintenant $\mathcal{J} = \mathcal{J}^- \cup \mathcal{J}^0$, et $\mathcal{Y}$ un sous-ensemble de $\mathcal{J}$. Formons une matrice $A_{\mathcal{Y}}$ dont les lignes sont les vecteurs a_i, $i \in \mathcal{Y}$, de sorte que c'est une matrice $m \times n$, m étant le nombre d'éléments de $\mathcal{Y}$.

LEMME 1.1. *Si les vecteurs a_i, $i \in \mathcal{Y}$, sont linéairement indépendants, la matrice $A_{\mathcal{Y}} A_{\mathcal{Y}}^*$ est régulière.*

DÉMONSTRATION. Soit $y \in E^m$ un vecteur non nul tel que

$$A_{\mathcal{Y}} A_{\mathcal{Y}}^* y = 0 \tag{1.3}$$

Alors

$$y^* A_{\mathcal{Y}} A_{\mathcal{Y}}^* y = (A_{\mathcal{Y}}^* y)^* A_{\mathcal{Y}}^* y = (A_{\mathcal{Y}}^* y, A_{\mathcal{Y}}^* y) = \| A_{\mathcal{Y}}^* y \|^2 = 0,$$

i.e.

$$A_{\mathcal{Y}}^* y = 0. \tag{1.4}$$

Mais $A_{\mathcal{Y}}^* y$ est tout simplement une combinaison linéaire de vecteurs a_i, $i \in \mathcal{Y}$, à coefficients y^i, $i = 1, \ldots, m$, où y^i sont les composantes du vecteur y. Etant donné l'hypothèse d'indépendance linéaire des a_i, $i \in \mathcal{Y}$, $A_{\mathcal{Y}}^* y$ ne peut p as être nulle. Donc (1.4) et, partant, (1.3) dont elle est la conséquence, sont impossibles. Ainsi, la matrice $A_{\mathcal{Y}} A_{\mathcal{Y}}^*$ n'annule qu'un vecteur nul, d'où sa régularité.

Posons

$$P = A_{\mathcal{Y}}^* (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}}. \tag{1.5}$$

On constate sans peine les propriétés suivantes de l'opérateur P:

$$PP = P, \tag{1.6}$$

$$P^* = P, \tag{1.7}$$

$$P(I - P) = (I - P) P = 0. \tag{1.8}$$

L'opérateur P est un opérateur de projection orthogonale dans un sous-espace engendré par les vecteurs a_i, $i \in \mathcal{Y}$.

En effet, on a pour tout vecteur $x \in E^n$

$$x = Px + (I - P) x.$$

Ensuite,

$$(Px, (I - P) x) = (x, P^* (I - P) x) = 0$$

en vertu de (1.7), (1.8) si bien que Px et $(I - P) x$ sont les composantes du développement orthogonal du vecteur x. De plus

$$Px = A_{\mathcal{Y}}^* u = \sum_{i \in \mathcal{Y}} a_i u^i,$$

où le vecteur $u \in E^m$ de composantes u^i est défini par la formule $u = (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} x$.

L'expression obtenue du vecteur Px montre qu'il est entièrement dans un sous-espace formé par les vecteurs a_i, $i \in \mathcal{Y}$.

Notons maintenant que

$$A_{\mathcal{Y}} (I - P) = A_{\mathcal{Y}} - (A_{\mathcal{Y}} A_{\mathcal{Y}}^*) (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} = 0. \tag{1.9}$$

Aussi, quel que soit $x \in E^n$, le vecteur $y = (I - P) x$ vérifie-t-il le système d'équations $A_{\mathcal{Y}} y = 0$.

2. Minimisation d'une fonction quadratique sur un sous-espace. Admettons qu'on a à minimiser la fonction quadratique $f(x)$ donnée par (1.1) avec les contraintes

$$(a_i, x) - b_i = 0, \ i \in \mathcal{Y}. \tag{1.10}$$

Les vecteurs a_i, $i \in \mathcal{Y}$, seront supposés linéairement indépendants.

Soit x_0 un point vérifiant (1.10).

Notons que si on désigne par $b_{\mathcal{Y}}$ le vecteur de composantes b_i, $i \in \mathcal{Y}$, le système (1.10) s'écrit $A_{\mathcal{Y}} x - b_{\mathcal{Y}} = 0$ de sorte que $A_{\mathcal{Y}} x_0 - b_{\mathcal{Y}} = 0$. Introduisons une nouvelle variable y par la formule

$$x = x_0 + (I - P) y \tag{1.11}$$

et considérons la fonction quadratique

$$\varphi(y) = f(x_0 + (I - P) y).$$

Comme le veulent les règles de dérivation d'une fonction composée et vu la symétrie de l'opérateur P, il y a entre les gradients de $\varphi(y)$ et $f(x)$ la relation

$$\varphi'(y) = (I - P) f'(x), \tag{1.12}$$

avec x et y liés par (1.11).

Lemme 1.2. *Soit $\bar{y}$ un point de minimum libre de la fonction $\varphi(y)$. Le point correspondant*

$$\bar{x} = x_0 + (I - P) \bar{y}$$

réalise alors le minimum de $f(x)$ sous les contraintes (1.10).

Démonstration. En $\bar{y}$ le gradient de $\varphi(y)$ s'annule : $\varphi'(\bar{y}) = 0$. Par suite, on a en vertu de (1.12)

$$(I - P) f'(\bar{x}) = 0,$$

ou

$$f'(\bar{x}) - A_{\mathcal{Y}}^* (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} f'(\bar{x}) = 0.$$

Posons $u = -(A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} f'(\bar{x})$, il vient

$$f'(x) + A_{\mathcal{Y}}^* u = 0. \tag{1.13}$$

On a de plus en utilisant (1.9)

$$A_y\bar{x} = A_y x_0 + A_y (I - P)\bar{y} = A_y x_0 = b_y,$$

i.e. $\bar{x}$ vérifie les conditions (1.10).

Ainsi, $\bar{x}$ est un point admissible et on a en ce point les conditions (1.13) qui sont les conditions nécessaires et suffisantes pour que $\bar{x}$ réalise le minimum de $f(x)$ sous (1.10). Le lemme est démontré.

Le lemme 1.2 montre qu'on peut ramener le problème de ce numéro à la minimisation de la fonction quadratique $\varphi(y)$ en l'absence de contraintes. La recherche du minimum de $\varphi(y)$ se fera par la méthode du gradient conjugué (ch. II, § 4) :

$$y_0 = 0, \quad p_1 = -\varphi'(0),$$

$$y_{k+1} = y_k + \alpha_{k+1} p_{k+1},$$

$$p_{k+1} = -\varphi'(y_k) + \frac{\|\varphi'(y_k)\|^2}{\|\varphi'(y_{k-1})\|^2} p_k.$$

Ici α_{k+1} se calcule selon la règle

$$\alpha_{k+1} = -\frac{(\varphi'(y_k), p_{k+1})}{(p_{k+1}, (I-P)C(I-P)p_{k+1})},$$

car la matrice définissant le terme quadratique de $\varphi(y)$ est de la forme (cela se vérifie aisément) :

$$(I - P)C(I - P).$$

Les formules ci-dessus définissent le processus par rapport aux variables auxiliaires y. Il y a cependant intérêt à passer en variables de départ x. Montrons au préalable qu'on a la relation

$$(I - P)p_k = p_k. \tag{1.14}$$

En effet, pour $k = 1$

$$(I-P)p_1 = -(I-P)\varphi'(0) = -(I-P)(I-P)f'(x_0) =$$
$$= -(I-P)f'(x_0) = -\varphi'(0) = p_1,$$

où nous nous sommes servis de (1.12) et du fait que

$$(I - P)(I - P) = I - P - (I - P)P = I - P.$$

Supposons la relation (1.14) vraie pour k et démontrons-la pour $k + 1$:

$$(I-P)p_{k+1} = -(I-P)\varphi'(y_k) + \frac{\|\varphi'(y_k)\|^2}{\|\varphi'(y_{k-1})\|^2}(I-P)p_k =$$
$$= -(I-P)(I-P)f'(x_k) + \frac{\|\varphi'(y_k)\|^2}{\|\varphi'(y_{k-1})\|^2} p_k =$$
$$= -(I-P)f'(x_k) + \frac{\|\varphi'(y_k)\|^2}{\|\varphi'(y_{k-1})\|^2} p_k = p_{k+1},$$

où nous avons une fois de plus utilisé (1.12) et (1.14).

La formule (1.11) entraîne maintenant

$$x_{k+1} = x_0 + (I - P)\, y_{k+1},$$

$$x_{k+1} = x_k + (I - P)(y_{k+1} - y_k) = x_k + (I - P)\, \alpha_{k+1} p_{k+1},$$

i.e.

$$x_{k+1} = x_k + \alpha_{k+1} p_{k+1}.$$

La prise en compte de (1.12) permet de transformer la formule donnant p_{k+1}:

$$p_{k+1} = -(I - P)\, f'(x_k) + \frac{\|(I-P)\, f'(x_k)\|^2}{\|(I-P)\, f'(x_{k-1})\|^2}\, p_k.$$

La formule pour α_{k+1} se récrit comme suit:

$$\alpha_{k+1} = -\frac{((I-P)\, f'(x_k),\ p_{k+1})}{((I-P)\, p_{k+1},\ C(I-P)\, p_{k+1})} = -\frac{(f'(x_k),\ p_{k+1})}{(p_{k+1},\ Cp_{k+1})}.$$

THÉORÈME 1.1. *Connaissant un point x_0 vérifiant* (1.10) *le problème de minimisation d'une fonction quadratique $f(x)$ avec les contraintes* (1.10) *est résolu en un nombre fini de pas par le processus*

$$p_1 = -(I - P)\, f'(x_0),$$

$$x_{k+1} = x_k + \alpha_{k+1} p_{k+1},$$

$$p_{k+1} = -(I - P)\, f'(x_k) + \frac{\|(I-P)\, f'(x_k)\|^2}{\|(I-P)\, f'(x_{k-1})\|^2}\, p_k,$$

$$\alpha_{k+1} = -\frac{(f'(x_k),\ p_{k+1})}{(p_{k+1},\ Cp_{k+1})}, \quad k = 0, 1, \ldots$$

La démonstration a été en fait effectuée plus haut lorsque nous avons déduit les formules du processus.

REMARQUE. On sait (ch. II, § 4) que si la méthode du gradient conjugué s'applique à une fonction quadratique de matrice C singulière, le processus converge en $n - l$ pas au plus, où l est le nombre de valeurs propres nulles de C. En recherchant le minimum de $\varphi(y)$ nous avons eu affaire à une fonction de matrice $(I - P)\, C \times \times (I - P)$. Mais comme $A_{\mathcal{Y}}(I - P) = 0$, i.e. $(I - P)\, A_{\mathcal{Y}}^* = 0$, on a $(I - P)\, a_i = 0$, $i \in \mathcal{Y}$. Dans le cas considéré la multiplicité de la valeur propre nulle de $(I - P)\, C\, (I - P)$ est donc au moins m, m étant le nombre des a_i, $i \in \mathcal{Y}$. C'est pourquoi le processus proposé ou bien converge vers le point de minimum ou bien montre l'absence de borne inférieure de $f(x)$ dans les conditions (1.10) au bout de $n - m$ pas au plus.

3. Algorithme pour le problème général de la programmation quadratique. Reprenons le problème général (1.1), (1.2). Posons

$$\mathcal{Y}(x) = \{i:\ (a_i,\ x) - b_i = 0,\ i \in \mathcal{J}^- \cup \mathcal{J}^0\}$$

pour tout point x vérifiant (1.2).

Nous supposerons remplie la condition de non-dégénérescence: quel que soit x, les vecteurs a_i, $i \in \mathcal{Y}(x)$, sont linéairement indépendants.

Décrivons l'algorithme de résolution du problème.

Soit x_0 un point quelconque vérifiant (1.2) et constituant l'approximation initiale. Prenons l'ensemble d'indices $\mathcal{Y}_0 = \mathcal{Y}(x_0)$ et construisons l'opérateur $P_{\mathcal{Y}_0}$:

$$P_{\mathcal{Y}_0} = A^*_{\mathcal{Y}_0}(A_{\mathcal{Y}_0}A^*_{\mathcal{Y}_0})^{-1}A_{\mathcal{Y}_0}.$$

Calculons les quantités

$$u_0 = -(A_{\mathcal{Y}_0}A^*_{\mathcal{Y}_0})^{-1}A_{\mathcal{Y}_0}f'(x_0), \quad (I - P_{\mathcal{Y}_0})f'(x_0) = f'(x_0) + A^*_{\mathcal{Y}_0}u_0.$$

Deux cas peuvent se présenter.

1) $(I - P_{\mathcal{Y}_0})f'(x_0) = 0$. Alors

$$f'(x_0) + A^*_{\mathcal{Y}_0}u_0 = 0, \tag{1.15}$$

et x_0 réalise le minimum de $f(x)$ sur une face définie par le système d'équations

$$(a_i, x) - b_i = 0, \quad i \in \mathcal{Y}_0$$

(voir ch. I, § 3).

Si le vecteur u_0^i n'a pas de composantes u_0, $i \in \mathcal{Y}(x_0) \cap \mathcal{J}^-$, négatives, le point x_0 est solution du problème originel (1.1), (1.2) car (1.15) sont alors les conditions nécessaires et suffisantes d'existence d'un minimum de $f(x)$ avec les contraintes (1.2).

Admettons qu'il existe un indice $j \in \mathcal{Y}(x_0) \cap \mathcal{J}^-$ tel que $u_0^j < 0$. Construisons un nouvel ensemble d'indices $\mathcal{Y}_0'$ en éliminant de $\mathcal{Y}_0$ l'indice j. Utilisons la méthode du gradient conjugué décrite au n° 2 pour minimiser $f(x)$ sous les contraintes

$$(a_i, x) - b_i = 0, \qquad i \in \mathcal{Y}_0'. \tag{1.16}$$

Le processus en question ne doit cependant pas quitter le domaine admissible (1.2). A chaque pas on effectue donc une vérification. On calcule

$$\bar{\alpha}_{k+1} = \min_i \frac{b_i - (a_i, x_k)}{(a_i, p_{k+1})}, \tag{1.17}$$

où on minimise par rapport à tous les i pour lesquels $(a_i, p_{k+1}) > 0$. Ici x_k est un point de la suite générée par l'algorithme et p_{k+1} la direction conjuguée en ce point. Soit maintenant α_{k+1} la grandeur correspondante du pas de la méthode du gradient conjugué. Si $\alpha_{k+1} < \bar{\alpha}_{k+1}$, alors $x_{k+1} = x_k + \alpha_{k+1}p_{k+1}$ et le processus continue. Si $\alpha_{k+1} \geqslant \bar{\alpha}_{k+1}$, alors $x_{k+1} = x_k + \bar{\alpha}_{k+1}p_{k+1}$ et l'algorithme s'arrête.

Conclusion: ou bien on réalise le minimum de $f(x)$ dans les conditions (1.16) ou bien le processus s'arrête quand $\alpha_{k+1} \geqslant \bar{\alpha}_{k+1}$.

Dans les deux cas on initialise avec le point obtenu comme on l'a fait avec le point initial x_0.

2) $(I - P_{\mathcal{Y}_0}) f'(x_0) \neq 0$.

Dans ce cas on résout par la technique du gradient conjugué le problème de minimisation de $f(x)$ sous les contraintes

$$(a_i, x) - b_i = 0, \ i \in \mathcal{Y}_0, \tag{1.18}$$

avec au départ le point x_0. Comme plus haut on teste à chaque pas l'admissibilité des points obtenus, i.e. on calcule chaque fois $\bar{\alpha}_{k+1}$ par les formules (1.17) et on applique l'algorithme tant que ou bien on ne réalise pas le minimum de $f(x)$ dans les conditions (1.18) ou bien on ne vérifie pas la condition $\alpha_{k+1} \geqslant \bar{\alpha}_{k+1}$ et on n'obtient pas le point $x_{k+1} = x_k + \bar{\alpha}_{k+1} p_{k+1}$. Dans les deux cas on recommence avec ce point le même processus qu'avec x_0.

Justifions la convergence de la méthode en un nombre fini de pas et montrons tout d'abord que dans le cas 1) comme dans 2) on fait un bon pas, i.e. on se déplace de x_0 en un point en lequel la valeur de $f(x)$ est strictement inférieure à $f(x_0)$. Du moment que les points sont générés par la méthode du gradient conjugué et que dans cette méthode la fonction décroît à chaque pas, la seule chose qu'il nous faut démontrer est que toujours $\bar{\alpha}_{k+1} > 0$, i.e. les contraintes (1.2) permettent un pas non nul suivant la direction choisie p_{k+1}, et que dans 1) x_0 n'est pas un point de minimum de $f(x)$ avec les contraintes (1.16) car, s'il n'en était pas ainsi, on ne se déplacerait pas du point x_0.

Démontrons deux lemmes auxiliaires.

Lemme 1.3. *Le vecteur* $p_1 = -(I - P_{\mathcal{Y}_0}) f'(x_0)$ *est solution du problème : minimiser*

$$\varphi(p) = (f'(x_0), p) + \frac{1}{2} \| p \|^2$$

avec les contraintes

$$A_{\mathcal{Y}_0} p = 0. \tag{1.19}$$

Démonstration. En effet, en vertu de (1.9) p_1 vérifie (1.19). De plus $\varphi'(p) = p + f'(x_0)$. Aussi

$$\varphi'(p_1) = p_1 + f'(x_0) = -(I - P_{\mathcal{Y}_0}) f'(x_0) + f'(x_0) =$$
$$= P_{\mathcal{Y}_0} f'(x_0) = -A^*_{\mathcal{Y}_0} u_0.$$

Donc

$$\varphi'(p_1) + A^*_{\mathcal{Y}_0} u_0 = 0, \tag{1.20}$$

ce qui est une condition nécessaire et suffisante pour que la fonction convexe $\varphi(p)$ atteigne son minimum en p_1 sous les contraintes (1.19), c.q.f.d.

Construisons le dual du problème de minimisation de $\varphi(p)$ avec les contraintes (1.19). Conformément aux règles du § 3, ch. I on doit trouver pour cela le minimum de la fonction $\varphi(p) + u^* A_{\mathcal{J}_0} p$. Dérivons par rapport à p et égalons les dérivées à zéro, il vient $p + f'(x_0) + A^*_{\mathcal{J}_0} u = 0$, i.e.

$$p = -f'(x_0) - A^*_{\mathcal{J}_0} u.$$

Par substitution de cette expression de p on a

$$\min_p \{\varphi(p) + u^* A_{\mathcal{J}_0} p\} = -\frac{1}{2} \| f'(x_0) + A^*_{\mathcal{J}_0} u \|^2.$$

Ainsi, le problème dual consiste à trouver le minimum par rapport à tous les vecteurs u de la fonction

$$\varphi^*(u) = -\frac{1}{2} \| f'(x_0) + A^*_{\mathcal{J}_0} u \|^2.$$

On vérifie aisément, en dérivant directement $\varphi^*(u)$ et en égalant à zéro les dérivées, que le vecteur

$$u_0 = -(A_{\mathcal{J}_0} A^*_{\mathcal{J}_0})^{-1} A_{\mathcal{J}_0} f'(x_0)$$

est solution du problème dual, i.e. il maximise $\varphi^*(u)$. Rappelons les composantes de u_0 : ce sont u_0^i, $i \in \mathcal{J}_0$. Ainsi, le vecteur u_0 est le vecteur des multiplicateurs de Lagrange dans le problème de minimisation de $\varphi(p)$ avec les contraintes (1.19). On constate de plus que la valeur commune (aux termes des théorèmes de dualité) du minimum de $\varphi(p)$ avec les contraintes (1.19) et du maximum par rapport à u de $\varphi^*(u)$ vaut

$$-\frac{1}{2} \| f'(x_0) + A^*_{\mathcal{J}_0} u_0 \|^2 \quad \text{ou} \quad -\frac{1}{2} \| (I - P_{\mathcal{J}_0}) f'(x_0) \|^2.$$

Lemme 1.4. *Etant donné une matrice $A_{\mathcal{J}_0'}$ formée en biffant dans $A_{\mathcal{J}_0}$ la ligne d'indice j tel que $u_0^j < 0$, et $(I - P_{\mathcal{J}_0}) f'(x_0) = 0$, le vecteur $\bar{p}_1 = -(I - P_{\mathcal{J}_0'}) f'(x_0)$ est différent de zéro et $(a_j, \bar{p}_1) < 0$.*

Démonstration. Le vecteur $\bar{p}_1$ s'écrit comme

$$\bar{p}_1 = -(f'(x_0) + A^*_{\mathcal{J}_0'} v), \quad v = -(A_{\mathcal{J}_0'} A^*_{\mathcal{J}_0'})^{-1} A_{\mathcal{J}_0'} f'(x_0).$$

Si $\bar{p}_1 = 0$, alors $f'(x_0) + A_{\mathcal{J}_0'} v = 0$. D'autre part,

$$(I - P_{\mathcal{J}_0}) f'(x_0) = f'(x_0) + A^*_{\mathcal{J}_0} u_0 = 0 \tag{1.21}$$

par hypothèse. On obtient après soustraction

$$A^*_{\mathcal{J}_0} u_0 - A^*_{\mathcal{J}_0'} v = u_0^j a_j + \sum_{i \neq j} (u_0^i - v^i) a_i = 0,$$

ce qui, étant donné $u_0^j \neq 0$, est impossible en vertu de l'indépendance linéaire des vecteurs a_i, $i \in \mathcal{J}_0$.

Démontrons la seconde partie du lemme.

Récrivons (1.21) en coordonnées:

$$f'(x_0) + \sum_{i \neq j} u_0^i a_i + (-u_0^j)(-a_j) = 0. \tag{1.22}$$

Notons que $-u_0^j > 0$ puisque $u_0^j < 0$.

Considérons le problème: minimiser $\varphi(p) = (p, f'(x_0)) + 1/2 \| p \|^2$ sous les contraintes

$$(a_i, p) = 0, \ i \in \mathcal{Y}_0', \tag{1.23}$$

$$-(a_j, p) \leqslant 0.$$

Du moment que $\varphi'(p) = f'(x_0) + p$, on a $\varphi'(0) = f'(x_0)$ et (1.22) est donc une condition nécessaire et suffisante pour que le point $p = 0$ réalise le minimum de $\varphi(p)$ avec les contraintes (1.23).

Par ailleurs, $\bar{p}_1$ est, selon le lemme 1.3, solution du problème de minimisation de $\varphi(p)$ en présence de contraintes $A_{\mathcal{Y}_0'} p = 0$, ou, en coordonnées,

$$(a_i, p) = 0, \ i \in \mathcal{Y}_0'. \tag{1.24}$$

Admettons que $(a_j, \bar{p}_1) \geqslant 0$. Le vecteur $\bar{p}_1$ vérifiant les contraintes (1.24), il vérifie alors toutes les contraintes (1.23). Mais

$$\begin{aligned}(f'(x_0), \bar{p}_1) &= -(f'(x_0), (I - P_{\mathcal{Y}_0'}) f'(x_0)) = \\ &= -(P_{\mathcal{Y}_0'} f'(x_0) + (I - P_{\mathcal{Y}_0'}) f'(x_0), (I - P_{\mathcal{Y}_0'}) f'(x_0)) = \\ &= -((I - P_{\mathcal{Y}_0'}) f'(x_0), (I - P_{\mathcal{Y}_0'}) f'(x_0)) = -\| \bar{p}_1 \|^2.\end{aligned}$$

Donc

$$\varphi(\bar{p}_1) = (f'(x_0), \bar{p}_1) + \frac{1}{2} \| \bar{p}_1 \|^2 = -\frac{1}{2} \| \bar{p}_1 \|^2 < 0,$$

ce qui contredit le fait que, dans les contraintes (1.23), la valeur minimale de $\varphi(p)$ est atteinte pour $p = 0$ et vaut 0. La contradiction ainsi obtenue montre que $(a_j, \bar{p}_1) < 0$ et le lemme se trouve démontré.

Voyons maintenant l'algorithme construit. Plaçons-nous dans le cas 1) et supposons que x_0 n'est pas solution du problème de programmation quadratique. On applique alors, selon l'algorithme, la méthode du gradient conjugué pour minimiser $f(x)$ avec les contraintes (1.16). Conformément aux formules de la méthode le premier pas est fait dans la direction du vecteur

$$\bar{p}_1 = -(I - P_{\mathcal{Y}_0'}) f'(x_0).$$

En vertu du lemme 1.4, $\bar{p}_1 \neq 0$ si bien que le point x_0 n'est pas solution du problème auxiliaire de minimisation que nous considérons.

Montrons que $\bar{\alpha}_1 > 0$. En effet, le vecteur $\bar{p}_1$ remplit la condition (1.24) et $(a_j, \bar{p}_1) < 0$ conformément au lemme 1.4. C'est pourquoi

$$(a_i, \bar{p}_1) \leqslant 0, \ i \in \mathcal{Y}_0. \tag{1.25}$$

Si $i \notin \mathcal{Y}_0$ on a d'après le choix de l'ensemble $\mathcal{Y}_0$

$$(a_i, x_0) - b_i < 0.$$

Donc

$$\bar{\alpha}_1 = \min_i \frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)} > 0,$$

car on ne minimise que par rapport à des i tels que $(a_i, \bar{p}_1) > 0$ ou, en fait, sur un sous-ensemble des i qui est disjoint, en vertu de (1.25), de $\mathcal{Y}_0$. Or, pour ces i on a $b_i - (a_i, x_0) > 0$.

L'inégalité $\bar{\alpha}_1 > 0$ montre que tous les points $x_0 + \alpha p_1$ satisfont pour $0 \leqslant \alpha \leqslant \bar{\alpha}_1$ aux conditions (1.2). En effet, pour $i \in \mathcal{Y}_0$

$$(a_i, x_0 + \alpha\bar{p}_1) - b_i = (a_i, x_0) - b_i + \alpha (a_i, \bar{p}_1) =$$
$$= \alpha (a_i, \bar{p}_1) = \begin{cases} = 0, & i \in \mathcal{Y}_0', \\ < 0, & i = j, \end{cases}$$

et pour $i \notin \mathcal{Y}_0$

$$(a_i, x_0 + \alpha\bar{p}_1) - b_i = (a_i, x_0) - b_i + \alpha (a_i, \bar{p}_1) < 0,$$

si $(a_i, p_1) \leqslant 0$. Si $(a_i, \bar{p}_1) > 0$, alors

$$\alpha \leqslant \bar{\alpha}_1 \leqslant \frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)}$$

et donc

$$(a_i, x_0 + \alpha\bar{p}_1) - b_i \leqslant (a_i, x_0) - b_i + \frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)} (a_i, \bar{p}_1) = 0.$$

Notons que la dernière inégalité est stricte si $\alpha < \bar{\alpha}_1$ ou $\bar{\alpha}_1 < \frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)}$.

Nous savons que deux cas peuvent se présenter: $\alpha_1 < \bar{\alpha}_1$ et $\alpha_1 \geqslant \bar{\alpha}_1$. Dans le premier cas, on obtient un nouveau point $x_1 = x_0 + \alpha_1 p_1$ qui vérifie les relations

$$(a_i, x_1) - b_i = 0, \ i \in \mathcal{Y}_0', \ (a_i, x_1) - b_i < 0, \ i \notin \mathcal{Y}_0', \tag{1.26}$$

et dans le second $x_1 = x_0 + \bar{\alpha}_1 p_1$ qu'on accepte pour un nouveau point initial où l'algorithme démarre en testant 1) ou 2). Ceci étant, x_1 vérifie, en plus des relations $(a_i, x_1) - b_i = 0$, $i \in \mathcal{Y}_0'$, les égalités $(a_i, x_1) - b_i = 0$ pour tous les $i \notin \mathcal{Y}_0'$, tels que $\frac{b_i - (a_i, x_0)}{(a_i, \bar{p}_1)} = \bar{\alpha}_1$.

Ainsi $\mathcal{Y}(x_1) \supset \mathcal{Y}_0'$, et l'inclusion est stricte.

Reprenons le cas $\alpha_1 < \bar{\alpha}_1$. L'algorithme de gradient conjugué continue alors, et tant que $\alpha_{k+1} < \bar{\alpha}_{k+1}$, tous les points x_{k+1} satisfont aux relations (1.26) comme x_1 parce qu'en vertu de la formule (1.14) et (1.19) avec $P = P_{\mathcal{Y}_0'}$

$$A_{\mathcal{Y}_0'} p_k = A_{\mathcal{Y}_0'} (I - P_{\mathcal{Y}_0'}) p_k = 0,$$

i.e. en coordonnées

$$(a_i, p_k) = 0, \ i \in \mathcal{Y}_0'.$$

Les inégalités

$$(a_i, x_k) - b_i < 0, \ i \notin \mathcal{Y}_0',$$

restent vraies elles aussi car, s'il n'en était pas ainsi, on serait dans le cas $\alpha_{k+1} \geqslant \bar{\alpha}_{k+1}$.

Ainsi, nous avons montré que dans le cas 1) le processus itératif génère la suite de points $x_0, x_1, \ldots, x_k$, $k \geqslant 1$, et que $f(x)$ est strictement décroissante suivant cette suite du moment que celle-ci est formée par la méthode des directions conjuguées. Le point terminal x_k ou bien réalise le minimum de $f(x)$ avec les contraintes (1.16) ou bien est un point en lequel $\mathcal{Y}_0'$ est strictement intérieur à $\mathcal{Y}(x_k)$.

Dans le cas 2) la direction de déplacement à partir du point x_0 coïncide avec le vecteur $p_1 = -(I - P_{\mathcal{Y}_0}) f'(x_0) \neq 0$, $(a_i, p_1) = 0$, $i \in \mathcal{Y}_0 (= (\mathcal{Y}(x_0))$, et on démontre donc sans peine que $\bar{\alpha}_1 > 0$ et que la méthode des directions conjuguées permet de faire au moins un pas non nul vers le point x_1 en lequel la valeur de $f(x)$ est strictement inférieure.

Toutes les démonstrations imitent celles que nous venons de faire. On aboutit à une suite de points $x_0, x_1, \ldots, x_k$, $k \geqslant 1$, x_k étant ou bien le point de minimum de $f(x)$ dans les conditions (1.18) ou bien tel que $\mathcal{Y}(x_k) \supset \mathcal{Y}_0$.

Si x_k réalise le minimum de $f(x)$ sous (1.16), il rend minimale $f(x)$ dans les deux cas sur une face définie par les relations

$$(a_i, x) - b_i = 0, \ i \in \mathcal{Y}(x_k), \tag{1.27}$$

car par construction $\mathcal{Y}(x_k) \supset \mathcal{Y}_0'$ dans le cas 1) et $\mathcal{Y}(x_k) \supset \mathcal{Y}_0$ dans le cas 2) et un point qui réalise le minimum sur un ensemble plus vaste le réalise également sur un autre qui est plus restreint.

Montrons qu'avec au départ un point x_0 on aboutit après un nombre fini de pas à x_k qui réalise lui-même le minimum de $f(x)$ sous les contraintes (1.27). En effet, l'exposé précédent autorise à dire que si la méthode du gradient conjugué ne résout pas le problème de minimisation, cela entraîne de suite une extension de l'ensemble des indices i tels qu'un point généré vérifie les relations $(a_i, x_k) - b_i = 0$. Les vecteurs a_i, $i \in \mathcal{Y}(x_k)$, étant linéairement

indépendants par hypothèse, cette extension cesse évidemment au bout de n pas au plus, où n est la dimension de x.

Ainsi l'algorithme proposé construit en n pas au plus un point x_k qui réalise le minimum de $f(x)$ sous les contraintes (1.27).

Notons que les ensembles $\mathcal{Y}(x_k)$ diffèrent pour ces x_k différents parce que la valeur de $f(x)$ décroît monotonement suivant la suite construite. En effet, soient x_m et x_k, $m < k$, deux points de minimum de $f(x)$ avec les contraintes

$$(a_i, x) - b_i = 0,$$

$i \in \mathcal{Y}(x_m)$ et $i \in \mathcal{Y}(x_k)$ respectivement. Si $\mathcal{Y}(x_m) = \mathcal{Y}(x_k)$, il est clair que $f(x_m) = f(x_k)$. Mais par construction $f(x_m) < f(x_k)$ pour $m > k$, d'où l'impossibilité de l'égalité $\mathcal{Y}(x_m) = \mathcal{Y}(x_k)$.

D'autre part, tous les ensembles $\mathcal{Y}(x_k)$ font partie de l'ensemble $\mathcal{J} = \mathcal{J}^- \cup \mathcal{J}^0$ et ils sont donc en nombre fini. Il s'ensuit l'arrêt du processus proposé après un nombre fini de pas. Or, l'arrêt ne peut avoir lieu que si l'on a trouvé le minimum de $f(x)$ sous les contraintes (1.2) car, dans le cas contraire, le processus peut continuer comme nous l'avons montré plus haut.

Cela signifie que le processus converge en un nombre fini de pas.

Remarque. Dans le cas de C singulière il peut arriver, selon la théorie des méthodes des directions conjuguées, que $(f'(x_k), p_{k+1}) \neq 0$ au point x_k, mais $(p_{k+1}, Cp_{k+1}) = 0$. Il est alors impossible de calculer α_{k+1} car

$$\alpha_{k+1} = -\frac{(f'(x_k), p_{k+1})}{(p_{k+1}, Cp_{k+1})}.$$

Puisque dans ce cas $f(x_k + \alpha p_{k+1})$ décroît indéfiniment avec la croissance de α, on peut poser $\alpha_{k+1} = +\infty$ et conduire les calculs comme d'ordinaire. Si $\alpha_{k+1} < +\infty$ l'algorithme de gradient conjugué aboutit au point $x_{k+1} = x_k + \alpha_{k+1} p_{k+1}$, ce qui ne contredit nullement les raisonnements ci-dessus. Si α_{k+1} n'est pas borné, i.e. $(a_i, p_{k+1}) \leqslant 0$ quel que soit i, le déplacement suivant la demi-droite $x_k + \alpha p_{k+1}$ entraîne une diminution indéfinie de la fonction $f(x)$. Ce fait signifie que le problème posé de la programmation quadratique n'admet pas de solution parce que la borne inférieure de $f(x)$ sous les contraintes (1.2) vaut $-\infty$.

4. Aspects numériques. L'algorithme proposé plus haut comprend en fait une seule opération compliquée, à savoir la projection du gradient sur un sous-espace, i.e. le calcul de $(I - P_{\mathcal{Y}}) f'(x)$. Ce calcul se fait de deux manières.

La première façon d'agir consiste à calculer directement la matrice $P_{\mathcal{Y}}$, i.e. $P_{\mathcal{Y}} = A_{\mathcal{Y}}^* (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}}$ auquel cas on a à calculer sans cesse la matrice $(A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1}$. Si celle-ci est connue, la recherche du

vecteur nécessaire $u = -(A_{\mathcal{Y}}A_{\mathcal{Y}}^*)^{-1}A_{\mathcal{Y}}f'(x)$ se ramène à multiplier la matrice par le vecteur.

Pour réduire le volume de calcul à chaque pas lorsque l'ensemble $\mathcal{Y}$ varie, on utilise le fait qu'en chassant de $\mathcal{Y}$ l'indice j la matrice $A_{\mathcal{Y}}A_{\mathcal{Y}}^*$ se trouve privée d'une ligne et d'une colonne, ce qui engendre la matrice $A_{\mathcal{Y}'}A_{\mathcal{Y}'}^*$. De même, si l'on adjoint à $\mathcal{Y}$ un indice, on voit $A_{\mathcal{Y}}A_{\mathcal{Y}}^*$ s'enrichir d'une colonne et d'une ligne supplémentaire. Cela permet d'utiliser les relations récurrentielles connues de l'algèbre linéaire (D. K. Faddeev, V. N. Faddeeva [57]). A savoir, soit B une matrice $n \times n$ symétrique qu'on peut mettre sous la forme

$$B = \begin{pmatrix} D & u \\ u^* & b \end{pmatrix},$$

où D est $(n-1) \times (n-1)$, u un $(n-1)$-vecteur-colonne, u^* son transposé, b un nombre. On vérifie aisément que

$$B^{-1} = \begin{pmatrix} D^{-1} + \dfrac{D^{-1}uu^*D^{-1}}{\alpha}, & -\dfrac{D^{-1}u}{\alpha} \\ -\dfrac{u^*D^{-1}}{\alpha}, & \dfrac{1}{\alpha} \end{pmatrix},$$

$$\alpha = b - u^*D^{-1}u.$$

Ainsi, connaissant D^{-1}, la matrice B^{-1} avec B obtenue en complétant la dernière colonne et la dernière ligne se calcule facilement.

Inversement, si B^{-1} s'écrit

$$B^{-1} = \begin{pmatrix} G & p \\ p^* & m \end{pmatrix},$$

on a pour D^{-1}

$$D^{-1} = G - \frac{pp^*}{m}.$$

Si la nouvelle matrice s'obtient donc de l'ancienne par suppression de la dernière ligne et de la dernière colonne ou, au contraire, par adjonction de la dernière ligne et de la dernière colonne, les matrices inverses sont trouvées moyennant de simples opérations arithmétiques. Le fait de biffer ou d'ajouter dans les formules écrites la dernière colonne et la dernière ligne ne joue aucun rôle car une permutation de lignes d'une matrice conduit, on le vérifie sans peine, à une permutation de colonnes de son inverse et une permutation de colonnes à celle de lignes.

Nous avons donc montré que la matrice de projection peut se calculer par des formules de récurrence. Le défaut de tels calculs est une accumulation éventuelle d'une grande erreur de calcul.

Et voici une autre façon d'effectuer les calculs.

D'après le n° 3, le vecteur $p_0 = -(I - P_{\mathcal{Y}}) f'(x)$ est solution du problème : minimiser $(f'(x), p) + \frac{1}{2} \| p \|^2$ avec les contraintes $A_{\mathcal{Y}} p = 0$. Il y a intérêt à passer au problème dual qui consiste, on l'a montré, en la maximisation de la fonction quadratique

$$-\frac{1}{2} \| f'(x) + A_{\mathcal{Y}}^* u \|^2$$

par rapport au vecteur u en l'absence de contraintes et qui se résout sans peine par la méthode des directions conjuguées. Sa solution est (voir n° 3) le vecteur $u_0 = -(A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} f'(x)$, i.e. le vecteur utilisé dans l'algorithme de résolution du problème général de programmation quadratique. Le vecteur p_0 en fonction de u_0 est obtenu facilement par la formule

$$p_0 = -(I - P_{\mathcal{Y}}) f'(x) = -[f'(x) - A_{\mathcal{Y}}^* (A_{\mathcal{Y}} A_{\mathcal{Y}}^*)^{-1} A_{\mathcal{Y}} f'(x)] =$$
$$= -[f'(x) + A_{\mathcal{Y}}^* u_0],$$

i.e.

$$p_0 = -[f'(x) + A_{\mathcal{Y}}^* u_0].$$

Ainsi, en calculant de cette façon, le problème se ramène à une application à plusieurs reprises de la procédure standard de la méthode des directions conjuguées.

5. Problème de programmation quadratique avec des contraintes simples. Le problème avec des contraintes simples est : minimiser

$$f(x) = \frac{1}{2} (x, Cx) + (d, x)$$

sous les contraintes $x^i \geqslant 0$, $i \in \mathcal{J}$, où $\mathcal{J}$ est un sous-ensemble de $\{1, 2, \ldots, n\}$. L'algorithme du n° 3 se simplifie alors considérablement. Au lieu de le simplifier formellement, formulons l'algorithme de résolution du problème proposé. Nous constaterons que la démonstration de sa convergence en un nombre fini de pas coïncide avec celle de l'algorithme du n° 3.

Soit donc x_0 un point arbitraire vérifiant les contraintes $x_0^i \geqslant 0$, $i \in \mathcal{J}$. Posons

$$\mathcal{J}(x) = \{i : x_i = 0,\ i \in \mathcal{J}\}.$$

Décrivons les opérations à une itération à partir de x_0. Calculons l'ensemble $\mathcal{Y}(x_0)$. Deux cas peuvent se présenter :

1) $(f'(x_0))^i = 0$, $i \notin \mathcal{Y}(x_0)$, où $(f'(x_0))^i$ est la i-ième composante du vecteur $f'(x_0)$.

Dans ce cas x_0 réalise le minimum de $f(x)$ dans les contraintes $x^i = 0$, $i \in \mathcal{Y}(x_0)$. Si de plus $(f'(x_0))^i \geqslant 0$ pour $i \in \mathcal{Y}(x_0)$, alors x_0 est solution du problème car on a en ce point les conditions nécessaires et suffisantes de minimum (voir ch. I, § 3).

Soit maintenant $(f'(x_0))^i < 0$ pour certains $i \in \mathcal{Y}(x_0)$. Posons

$$\mathcal{Y}' = \{i \in \mathcal{Y}(x_0) : (f'(x_0))^i \geqslant 0\}.$$

Trouvons le minimum de $f(x)$ par la méthode du gradient conjugué en prenant comme variables les seuls x^i, $i \notin \mathcal{Y}'$, et en laissant $x^i = 0$ pour $i \in \mathcal{Y}'$. Avec cette méthode on a à calculer toujours $\bar{\alpha}_{k+1}$:

$$\bar{\alpha}_{k+1} = \min_i \left(-\frac{x_k^i}{p_{k+1}^i} \right),$$

où le minimum est réalisé par rapport à tous les $i \notin \mathcal{Y}'$, tels que $p_{k+1}^i < 0$, et on compare α_{k+1} et $\bar{\alpha}_{k+1}$. Si $\alpha_{k+1} < \bar{\alpha}_{k+1}$, alors $x_{i+1}^i = x_i + \alpha_{k+1} p_{k+1}^i$, $i \notin \mathcal{Y}'$, $x_{k+1}^i = x_k^i = 0$, $i \in \mathcal{Y}'$. Si $\alpha_{k+1} \geqslant \bar{\alpha}_{k+1}$, alors $x_{k+1}^i = x_k^i + \bar{\alpha}_{k+1} p_{k+1}^i$, $i \notin \mathcal{Y}'$, $x_{k+1}^i = x_k^i = 0$, $i \in \mathcal{Y}'$. Le processus s'arrête après un nombre fini de pas et on trouve un point x_{k+1} tel que $f(x)$ atteint en ce point son minimum dans les conditions $x^i = 0$, $i \in \mathcal{Y}'$, ou tel que $\alpha_{k+1} \geqslant \bar{\alpha}_{k+1}$. Dans ce cas $\mathcal{Y}(x_k) \supset \mathcal{Y}'$, et l'inclusion est telle qu'il existe des $i \in \mathcal{Y}(x_k)$, mais $i \notin \mathcal{Y}'$. Dans l'un et dans l'autre cas le processus recommence avec x_{k+1} au départ.

2) Il existe des indices i tels que $(f'(x_0))^i \neq 0$, $i \notin \mathcal{Y}(x_0)$. On minimise alors $f(x)$ par rapport aux variables x^i, $i \notin \mathcal{Y}(x_0)$, par la méthode du gradient conjugué, les composantes x^i, $i \in \mathcal{Y}(x_0)$, étant toujours nulles. De plus, on calcule à chaque pas (comme pour 1)) la quantité

$$\bar{\alpha}_{k+1} = \min_i \left(-\frac{x_k^i}{p_{k+1}^i} \right),$$

où l'on minimise par rapport à tous les $i \notin \mathcal{Y}(x_0)$, $p_{k+1}^i < 0$. Le processus se termine comme dans le cas 1).

On voit aisément qu'en raisonnant comme au n° 3 on démontre la convergence de l'algorithme en un nombre fini de pas ou on établit l'absence de borne inférieure de $f(x)$ dans les conditions $x^i \geqslant 0$, $i \in \mathcal{J}$.

§ 2. Méthode des directions admissibles

Il s'agit d'une des premières méthodes de résolution du problème de programmation convexe.

On demande de minimiser $f_0(x)$ avec les contraintes

$$f_i(x) \leqslant 0, \quad i = 1, \ldots, m,$$
$$Ax - b = 0. \tag{2.1}$$

Ici $x \in E^n$, $f_i(x)$, $i = 0, 1, \ldots, m$, sont convexes et admettent des dérivées continues, A est une matrice $l \times m$ et b un l-vecteur.

Nous supposerons de plus que les gradients des fonctions $f_i(x)$, $i = 0, 1, \ldots, m$, satisfont à la condition de Lipschitz

$$\| f_i'(x_1) - f_i'(x_2) \| \leqslant C \| x_1 - x_2 \| \tag{2.2}$$

et $\| f_i'(x) \| \leqslant K$ pour tous les points x qui intreviendront dans la suite.

Désignons par D un domaine admissible, i.e. l'ensemble

$$D = \{x : f_i(x) \leqslant 0,\ i = 1, \ldots, m,\ Ax - b = 0\}.$$

L'ensemble D sera supposé compact si bien qu'on a la condition de borne des gradients. Soit x_0 un point de D. Trouvons une direction $p \in E^n$ telle que, pour α petits, $(x_0 + \alpha p) \in D$ et $f_0(x_0 + \alpha p) < f_0(x_0)$. Une telle direction est dite admissible. En faisant un pas α_1 dans cette direction on trouve un nouveau point $x_1 = x_0 + \alpha_1 p \in D$. Le processus recommence à partir de x_1. Nous devons maintenant élaborer une technique efficace de recherche de directions admissibles et de choix du pas α de façon à garantir la convergence vers le point de minimum.

La condition de non-dégénérescence suivante est supposée remplie partout dans la suite: il existe un point $\bar{x}$ tel que

$$A\bar{x} - b = 0,\ f_i(\bar{x}) < 0,\ i = 1, \ldots, m.$$

1. Méthode du choix d'une direction admissible. Posons

$$\mathcal{J}_\delta(x) = \{i : f_i(x) \geqslant -\delta,\ i = 1, \ldots, m\}$$

pour tout point $x \in D$. Soient $\xi_i > 0$, $i = 0, 1, \ldots, m$, des nombres quelconques. Considérons en tout $x \in D$ le problème suivant:

$$\begin{gathered} \min \eta, \\ (f_i'(x), p) \leqslant \xi_i \eta,\ i \in \mathcal{J}_\delta(x) \cup \{0\}, \\ Ap = 0,\ \| p \| \leqslant 1. \end{gathered} \tag{2.3}$$

Ici η est un nombre et $\| p \|$ une norme arbitraire. Pour que (2.3) soit un problème de programmation linéaire, il est commode de prendre pour norme

$$\| p \| = \max_{1 \leqslant i \leqslant n} | p^i |.$$

Soit $p_\delta(x)$, $\eta_\delta(x)$ une solution du problème (2.3). Comme le vecteur $p = 0$, $\eta = 0$, vérifie les contraintes (2.3), il est clair que $\eta_\delta(x) \leqslant 0$. Montrons que si $\eta_\delta(x) < 0$, alors $p_\delta(x)$ est une direction admissible.

En effet, soit $\alpha > 0$. On a pour $i = 0$ conformément à la formule de Taylor

$$f_0(x+\alpha p_\delta(x)) = f_0(x) + \alpha(f_0'(\theta_0),\ p_\delta(x)) = f_0(x) + \\ + \alpha(f_0'(x),\ p_\delta(x)) + \alpha(f_0(\theta_0) - f_0'(x),\ p_\delta(x)) \leqslant \\ \leqslant f_0(x) + \alpha(f_0'(x),\ p_\delta(x)) + \alpha^2 C \| p_\delta(x) \|^2,$$

où $\theta_0 = x + \zeta_0 \alpha p_\delta(x)$, $0 \leqslant \zeta_0 \leqslant 1$, et nous avons utilisé l'inégalité

$$\| f_0'(\theta_0) - f_0'(x) \| \leqslant C \| \theta_0 - x \| \leqslant C\alpha \| p_\delta(x) \|.$$

Ensuite, $(f_0(x),\ p_\delta(x)) \leqslant \xi_0 \eta_\delta(x)$ en vertu de (2.3). Donc

$$f_0(x+\alpha p_\delta(x)) \leqslant f_0(x) + \alpha \xi_0 \eta_\delta(x) + \alpha^2 C \| p_\delta(x) \|^2. \tag{2.4}$$

On a de même pour $i \in \mathcal{J}_\delta^-(x)$

$$f_i(x+\alpha p_\delta(x)) \leqslant f_i(x) + \alpha \xi_i \eta_\delta(x) + \alpha^2 C \| p_\delta(x) \|^2 \tag{2.5}$$

et pour $i \notin \mathcal{J}_\delta^-(x)$

$$f_i(x+\alpha p_\delta(x)) = f_i(x) + \alpha(f_i'(\theta_i),\ p_\delta(x)) \leqslant f_i(x) + \alpha K \| p_\delta(x) \|. \tag{2.6}$$

Choisissons maintenant un $\alpha > 0$ tel qu'on ait les inégalités

$$\begin{aligned} &f_0(x+\alpha p_\delta(x)) \leqslant f_0(x) + \tfrac{1}{2}\alpha \xi_0 \eta_\delta(x), \\ &f_i(x+\alpha p_\delta(x)) \leqslant 0, \quad i \in \mathcal{J}_\delta^-(x), \\ &f_i(x+\alpha p_\delta(x)) \leqslant 0, \quad i \notin \mathcal{J}_\delta^-(x). \end{aligned} \tag{2.7}$$

Il suffit pour cela d'avoir les inégalités

$$\begin{aligned} 1 + \frac{\alpha C \| p_\delta(x) \|^2}{\xi_0 \eta_\delta(x)} \leqslant \frac{1}{2}, \quad \xi_i \eta_\delta(x) + \alpha C \| p_\delta(x) \|^2 \leqslant 0, \quad i \in \mathcal{J}_\delta^-(x), \\ -\delta + \alpha K \| p_\delta(x) \| \leqslant 0, \quad i \notin \mathcal{J}_\delta^-(x), \end{aligned} \tag{2.8}$$

puisque, par suite de (2.4), (2.5),

$$f_0(x+\alpha p_\delta(x)) \leqslant f_0(x) + \alpha \xi_0 \eta_\delta(x) \left[1 + \frac{\alpha C \| p_\delta(x) \|^2}{\xi_0 \eta_\delta(x)} \right],$$

$$f_i(x+\alpha p_\delta(x)) \leqslant f_i(x) + \alpha [\xi_i \eta_\delta(x) + \alpha C \| p_\delta(x) \|^2], \quad i \in \mathcal{J}_\delta^-(x),$$

et comme $f_i(x) \leqslant -\delta$ pour $i \notin \mathcal{J}_\delta^-(x)$, on a

$$f_i(x + \alpha p_\delta(x)) \leqslant -\delta + \alpha K \| p_\delta(x) \|.$$

On obtient à partir de (2.8)

$$\alpha \leqslant -\frac{1}{2} \frac{\xi_0 \eta_\delta(x)}{C \| p_\delta(x) \|^2}, \quad \alpha \leqslant -\frac{\xi_i \eta_\delta(x)}{C \| p_\delta(x) \|^2}, \quad \alpha \leqslant \frac{\delta}{K \| p_\delta(x) \|}. \tag{2.9}$$

Si α vérifie les inégalités (2.9), on a donc les inégalités (2.7). Vu que $Ap_\delta(x) = 0$ et donc $A(x + \alpha p_\delta(x)) - b = Ax - b + \alpha A p_\delta(x) = 0$, il en résulte que $p_\delta(x)$ est effectivement une

direction admissible. Montrons à présent que si x ne coïncide pas avec la solution x_*, le point de minimum de $f_0(x)$ dans le domaine D, alors $\eta_\delta(x) < 0$ pour tous les δ suffisamment faibles.

LEMME 2.1. *Si $x \in D$ n'est pas solution du problème de minimisation de $f_0(x)$ avec les contraintes* (2.3), *alors $\eta_\delta(x) < 0$ pour tous les δ suffisamment petits.*

DÉMONSTRATION. Rappelons que nous supposons remplies les conditions de non-dégénérescence, i.e. il existe un point $\bar{x}$ tel que

$$A\bar{x} - b = 0, \; f_i(\bar{x}) \leqslant \sigma, \; i = 1, \ldots, m, \; \sigma < 0. \tag{2.10}$$

Soit x_* solution du problème. Si

$$\mathcal{J}_0^-(x) = \{i : f_i(x) = 0, \; i = 1, \ldots, m\},$$

alors pour $\delta < \delta_0$

$$-\delta_0 = \max_{i \notin \mathcal{J}_0^-(x)} f_i(x),$$

$\mathcal{J}_\delta^-(x) = \mathcal{J}_\delta^-(x)$. En effet, si $i \in \mathcal{J}_\delta^-(x)$, $\delta < \delta_0$, alors $f_i(x) \geqslant -\delta$. Mais pour tous les $i \notin \mathcal{J}_0^-(x)$, $f_i(x) \leqslant -\delta_0 < -\delta$, i.e. $i \in \mathcal{J}_0^-(x)$. Nous admettrons $\delta < \delta_0$ de sorte que $\mathcal{J}_\delta^-(x) = \mathcal{J}_0^-(x)$. Posons

$$x_\rho = \rho\bar{x} + (1-\rho)x_*, \quad 0 < \rho \leqslant 1.$$

Les fonctions $f_i(x)$, $i = 0, 1, \ldots, m$, étant convexes et $f_i(x_*) \leqslant 0$, $i = 1, \ldots, m$, on a alors

$$f_i(x_\rho) \leqslant \rho f_i(\bar{x}) + (1-\rho) f_i(x_*) \leqslant \rho\sigma, \quad i = 1, \ldots, m.$$

Ensuite, $f_i(x) = 0$ pour $i \in \mathcal{J}_\delta^-(x)$ et donc pour $0 < \lambda < 1$

$$\lambda\rho\sigma \geqslant \lambda f_i(x_\rho) = \lambda f_i(x_\rho) + (1-\lambda) f_i(x) \geqslant f_i(\lambda x_\rho + (1-\lambda)x) =$$
$$= f_i(x + \lambda(x_\rho - x)) - f_i(x) \geqslant \lambda(f_i'(x), x_\rho - x),$$

où nous avons utilisé l'inégalité (ch. I, § 2) vraie pour toute fonction convexe dérivable :

$$f(y) - f(x) \geqslant (f'(x), y - x).$$

Donc

$$\rho\sigma \geqslant (f_i'(x), x_\rho - x), \; i \in \mathcal{J}_\delta^-(x). \tag{2.11}$$

Puisque le point x ne réalise le minimum de $f(x)$ dans D, on a

$$0 > \gamma = f_0(x_*) - f_0(x) \geqslant (f_0'(x), x_* - x).$$

D'où

$$(f_0'(x), x_\rho - x) = \rho(f_0'(x), \bar{x} - x) + (1-\rho)(f_0'(x), x_* - x) \leqslant$$
$$\leqslant \rho(f_0'(x), \bar{x} - x) + (1-\rho)\gamma. \tag{2.12}$$

Il découle de (2.11) et (2.12) que, pour des $\rho > 0$ suffisamment petits,

$$(f_0'(x), \bar{p}_\rho) < 0, \ (f_i'(x), \bar{p}_\rho) < 0, \ i \in \mathcal{J}_\delta^-(x), \tag{2.13}$$

avec $\bar{p}_\rho = x_\rho - x$ et où on a tenu compte de $\sigma < 0$, $\gamma < 0$. Posons $p_\rho = \bar{p}_\rho$ si $\|\bar{p}_\rho\| \leqslant 1$ et $p_\rho = \frac{\bar{p}_\rho}{\|\bar{p}_\rho\|}$ si $\|\bar{p}_\rho\| > 1$ de sorte que $\|p_\rho\| \leqslant 1$. Posons de plus

$$\eta_\rho = \max_{i \in \mathcal{J}_\delta^-(x) \cup \{0\}} \frac{(f_i'(x), p_\rho)}{\xi_i}.$$

Par suite de (2.13), $\eta_\rho < 0$ et on a les inégalités

$$(f_i'(x), p_\rho) \leqslant \xi_i \eta_\rho, \ i \in \mathcal{J}_\delta^-(x) \cup \{0\}, \ \eta_\rho < 0. \tag{2.14}$$

Comme $x_\rho = \rho\bar{x} + (1 - \rho) x_*$ et

$$A\bar{x} - b = 0, \ Ax_* - b = 0,$$

on a

$$Ax_\rho - b = 0.$$

Notons que $p_\rho = \alpha\bar{p}_\rho = \alpha(x_\rho - x)$, où $0 < \alpha \leqslant 1$. Aussi

$$Ap_\rho + (Ax - b) = \alpha[Ax_\rho - b] + (1 - \alpha)[Ax - b] = 0. \tag{2.15}$$

Il résulte de (2.14) et (2.15) que le vecteur p_ρ et la quantité η_ρ vérifient les conditions (2.3). Du moment que $\eta_\rho < 0$, on a *a fortiori* $\eta_\delta(x) < 0$ car $\eta_\delta(x) \leqslant \eta_\rho$ par définition. Le lemme est démontré.

2. Algorithme de la méthode des directions admissibles. Soit $x_0 \in D$ une approximation initiale quelconque et $\delta_0 > 0$. Décrivons le pas général de l'algorithme. Admettons qu'au k-ième pas on a obtenu un point $x_k \in D$ et $\delta_k > 0$.

En résolvant le problème

$$\min \eta,$$
$$(f_i'(x_k), p) \leqslant \xi_i \eta, \quad i \in \mathcal{J}_\delta^-(x_k) \cup \{0\},$$
$$Ap = 0, \quad \|p\| \leqslant 1,$$

nous trouverons $p_{\delta_k}(x_k) \equiv p_k$ et $\eta_{\delta_k}(x_k) \equiv \eta_k$.

Remarque. Si l'on prend pour norme du vecteur p la quantité $\max_i |p^i|$, le problème ci-dessus est celui de programmation linéaire qui se prête à une résolution par une technique standard.

Deux cas sont possibles.

1) $\eta_k < -\delta_k$.

On pose successivement $\alpha = \frac{1}{2^i}$, $i = 0, 1, \ldots$, et on trouve le premier i_0 vérifiant les inégalités

$$f_0\left(x_k + \frac{1}{2^{i_0}} p_k\right) \leqslant f_0(x_k) + \frac{1}{2 \cdot 2^{i_0}} \xi_0 \eta_k,$$

$$f_i\left(x_k + \frac{1}{2^{i_0}} p_k\right) \leqslant 0, \quad i = 1, \ldots, m.$$

Posons $\alpha_k = \frac{1}{2^{i_0}}$ et $x_{k+1} = x_k + \alpha_k p_k$, $\delta_{k+1} = \delta_k$, de sorte que

$$\begin{aligned} &f_0(x_{k+1}) \leqslant f_0(x_k) + \frac{1}{2} \alpha_k \xi_0 \eta_k, \\ &f_i(x_{k+1}) \leqslant 0, \quad i = 1, \ldots, m. \end{aligned} \tag{2.16}$$

2) $\eta_k \geqslant -\delta_k$.

Posons $x_{k+1} = x_k$, $\delta_{k+1} = \frac{1}{2} \delta_k$.

Dans le premier cas on arrive donc en un nouveau point et dans le second ce déplacement n'a pas lieu.

Enonçons la condition d'arrêt de l'algorithme: si à un pas k on a $\delta_k < \delta^0(x_k)$, où

$$\delta^0(x_k) = -\max_{i \notin \mathcal{J}_0^-(x_k)} f_i(x_k)$$

et $\eta_k = 0$, alors x_k est solution du problème posé, i.e. c'est le point de minimum de $f_0(x)$ avec les contraintes (2.1).

3. Justification de la convergence de l'algorithme. Montrons que si la suite x_k s'arrête à un pas k parce que la condition que nous venons d'énoncer est remplie, x_k est effectivement solution du problème. En effet, admettons que nous avons $\eta_k = \eta_{\delta_k}(x_k) = 0$ et

$$\delta_k < \delta^0(x_k) = -\max_{i \notin \mathcal{J}_0^-(x_k)} f_i(x_k). \tag{2.17}$$

Or, nous avons constaté en démontrant le lemme 2.1 que, s'il y a (2.17), $\eta_{\delta_k}(x_k) < 0$ à condition que x_k ne soit pas solution du problème. Etant donné que $\eta_k = 0$ par hypothèse, il en résulte que x_k réalise le minimum de $f_0(x)$ pour $x \in D$. Supposons maintenant que le processus itératif se poursuit sans fin de sorte que nous avons une suite x_k infinie, $k = 0, 1, \ldots$ Soit x_k le point en lequel $\eta_k < -\delta_k$, i.e. on est dans le cas 1). En utilisant les estimations (2.9)

et $\| p_k \| = \| p_{\delta_k}(x_k) \| \leqslant 1$, on peut dire que si les inégalités

$$\begin{aligned} &\alpha \leqslant -\frac{1}{2}\frac{\xi_0\eta_k}{C}, \\ &\alpha \leqslant -\frac{\xi_i\eta_k}{C}, \quad i \in \mathcal{J}_k^-, \\ &\alpha \leqslant \frac{\delta_k}{K}, \quad i \notin \mathcal{J}_k^-, \end{aligned} \tag{2.18}$$

où $\mathcal{J}_k^- = \mathcal{J}_{\delta_k}^-(x_k)$ pour simplifier l'écriture, sont justes, on a (2.7):

$$f_0(x_k + \alpha p_k) \leqslant f_0(x_k) + \frac{1}{2}\alpha\xi_0\eta_k,$$

$$f_i(x_k + \alpha p_k) \leqslant 0, \quad i = 1, \ldots, m.$$

Rappelons que, conformément à l'algorithme, α_k coïncide avec la première quantité $1/2^i$, $i = 0, 1, \ldots$, qui vérifie les inégalités (2.16). Conséquence: les inégalités sont satisfaites après un nombre fini d'essais.

Soit i_0 le premier indice qui vérifie ces inégalités si bien que $\alpha = 1/2^{i_0}$. Cela signifie que (2.16) n'ont pas eu lieu pour $\alpha = 1/2^{i_0-1}$ et la quantité α n'a pas vérifié (2.18), i.e.

$$\frac{1}{2^{i_0-1}} > \min\left\{-\frac{1}{2}\frac{\xi_0\eta_k}{C}, \frac{\delta_k}{K}, \min_{i\in\mathcal{J}_k^-} -\frac{\xi_i\eta_k}{C}\right\}.$$

Donc

$$\alpha_k = \frac{1}{2^{i_0}} > \frac{1}{2}\min\left\{-\frac{\xi_0\eta_k}{2C}, \frac{\delta_k}{K}, \min_{i\in\mathcal{J}_k^-} -\frac{\xi_i\eta_k}{C}\right\}. \tag{2.19}$$

Vu que dans le cas considéré $-\eta_k > \delta_k$, on renforce (2.19) en remplaçant $-\eta_k$ par δ_k. On obtient de ce fait

$$\alpha_k > \frac{\delta_k}{2}\varepsilon_0, \quad \varepsilon_0 = \min\left\{\frac{\xi_0}{2C}, \frac{\xi_1}{C}, \frac{\xi_2}{C}, \ldots, \frac{\xi_m}{C}, \frac{1}{K}\right\}. \tag{2.20}$$

Utilisons (2.16), (2.20) et $\eta_k < -\delta_k < 0$, il vient

$$f_0(x_{k+1}) \leqslant f_0(x_k) + \frac{1}{2}\alpha_k\xi_0\eta_k \leqslant f_0(x_k) - \frac{\xi_0\varepsilon_0}{4}\delta_k^2. \tag{2.21}$$

Il en découle $\delta_k \to 0$ pour $k \to \infty$. En effet, la suite δ_k, $k = 0, 1, \ldots$, décroît monotonement, et, si $\delta_{k+1} < \delta_k$, alors $\delta_{k+1} = \frac{1}{2}\delta_k$, si bien que $\delta_k \not\to 0$ ne signifie qu'une seule chose, à savoir que $\delta_k = \delta > 0$ pour tous les k suffisamment grands. Mais si δ_k reste constant, on a la condition $\eta_k < -\delta_k$ et donc l'inégalité (2.21).

Ainsi, pour tous les k suffisamment importants ($k \geqslant k_0$), $\delta_k = \delta$ et

$$f_0(x_{k+1}) \leqslant f_0(x_k) - \frac{\xi_0\varepsilon_0}{4}\delta^2.$$

C'est pourquoi

$$f_0(x_N) \leqslant f_0(x_{k_0}) - (N - k_0)\frac{\xi_0\varepsilon_0}{4}\delta^2,$$

i.e. $f_0(x_N) \to -\infty$ quand $N \to \infty$, ce qui contredit la propriété d'une fonction continue $f_0(x)$ d'être bornée dans un domaine compact D.

Ainsi, $\delta_k \to 0$. Or, cela signifie que δ_0 initial est divisé par deux un nombre infini de fois, i.e. le cas 2) : $\eta_k \geqslant -\delta_k$, a lieu un nombre infini de fois.

Soit $\mathcal{Y}$ l'ensemble des indices k pour lesquels on a eu 2). On a alors $\eta_k \to 0$ lorsque $k \to \infty$, $k \in \mathcal{Y}$. Cela résulte de suite de l'inégalité $-\delta_k \leqslant \eta_k \leqslant 0$ et de la tendance vers zero de δ_k. Considérons une suite de points $x_k \in D$, $k \in \mathcal{Y}$. D étant compact, on estime, sans restreindre la généralité, que x_k converge vers un point x_*. Montrons que cette limite est le point de minimum de $f_0(x)$ dans D.

Supposons par l'absurde que x_* ne réalise pas le minimum de $f_0(x)$ dans D. On dit alors, d'après le lemme 2.1, que, quels que soient $\delta < \delta_0(x_*)$,

$$\delta_0(x_*) = -\max_{i \notin \mathcal{J}_0^-(x_*)} f_i(x_*),$$

$\mathcal{J}_\delta^-(x_*) = \mathcal{J}_0^-(x_*)$, et la quantité $\eta_\delta(x_*) < 0$. Il y plus: comme $\mathcal{J}_\delta^-(x_*) = \mathcal{J}_0^-(x_*)$, on a $\eta_\delta(x_*) = \eta_0(x_*) < 0$. Ensuite, $\mathcal{J}_{\delta_k}^-(x_k) \subseteq \mathcal{J}_0^-(x_*)$ pour des $k \in \mathcal{Y}$ suffisamment grands. En effet, soit $i \notin \mathcal{J}_0^-(x_*)$. Alors $f_i(x_*) < 0$. Par suite de $\delta_k \to 0$, on a donc $f_i(x_*) < -\delta_k$ pour k suffisamment élevés, et, comme $x_k \to x_*$, $f_i(x_k)$, k grand, est également inférieur à $-\delta_k$, i.e. $i \notin \mathcal{J}_{\delta_k}^-(x_k)$. Ainsi, si $i \notin \mathcal{J}_0^-(x_*)$, alors $i \notin \mathcal{J}_{\delta_k}^-(x_k)$ pour k importants, i.e. $\mathcal{J}_{\delta_k}^-(x_k) \subseteq \mathcal{J}_0^-(x_*)$. Du moment que par hypothèse x_* n'est pas un point de minimum de $f_0(x)$ dans D, il existe un vecteur $p(x_*)$ tel que $Ap(x_*) = 0$, $\| p(x_*) \| \leqslant 1$,

$$(f_i'(x_*),\ p(x_*)) \leqslant \xi_i \eta_0(x_*),\quad i \in \mathcal{J}_0^-(x_*) \cup \{0\},$$

de plus, comme on l'a déjà dit, $\eta_0(x_*) < 0$. Mais on a alors par continuité pour k grands

$$(f_i'(x_k),\ p(x_*)) \leqslant \frac{1}{2}\xi_i \eta_0(x_*),\quad i \in \mathcal{J}_k^- \cup \{0\},$$

$$Ap(x_*) = 0,\ \| p(x_*) \| \leqslant 1,$$

parce que $x_k \to x_*$, $\mathcal{J}_k^- \subseteq \mathcal{J}_0^-(x_*)$. Ces dernières relations signifient cependant que

$$\eta_k \equiv \eta_{\delta_k}(x_k) \leqslant \frac{1}{2}\eta_0(x_*) < 0$$

pour tous les k suffisamment grands, ce qui contredit $\eta_k \to 0$, $k \to +\infty$, $k \in \mathcal{Y}$, fait établi plus haut. La contradiction obtenue démontre que x_* réalise le minimum de $f_0(x)$ dans D.

THÉORÈME 2.1. *La suite de points x_k construite par la méthode des directions admissibles jouit de la propriété que, sans croître de façon monotone, $f_0(x_k)$ tend vers $f_0(x_*)$, où x_* est un point de minimum de $f_0(x)$ dans le domaine D.*

DÉMONSTRATION. Par construction de la suite $\{x_k\}$ $f_0(x_{k+1}) \leqslant \leqslant f_0(x_k)$ de sorte que la suite de nombres $f_0(x_k)$ n'est pas monotone croissante. Puisqu'elle est minorée, elle possède une limite $\bar{f}_0$. Nous avons cependant dit qu'il existe une suite partielle $\{x_k\}$, $k \in \mathcal{Y}$, telle que $x_k \to x_*$. Donc $f_0(x_k) \to f(x_*)$. Comme une suite convergente et sa sous-suite tendent vers une même limite il en résulte $f_0(x_k) \to f_0(x_*)$, c.q.f.d.

REMARQUE 1. Il y en a parmi les contraintes $f_i(x) \leqslant 0$ pour lesquelles les fonctions $f_i(x)$ sont linéaires. On montre sans peine, à force de développer un peu les raisonnements ci-dessus, qu'on peut poser $\xi_i = 0$ pour de tels indices i.

Il est possible de plus d'affaiblir la condition de non-dégénérescence, à savoir il suffit d'exiger l'existence d'un point $\bar{x} \in D$ tel que $f_i(\bar{x}) < 0$ dans le cas des seuls i pour lesquels $f_i(x)$ sont non linéaires.

REMARQUE 2. La suite x_k même peut en général ne pas converger, mais si x_* est l'unique point de minimum de $f_0(x)$ pour $x \in D$, on constate aisément que $x_k \to x_*$. La vitesse de convergence de la méthode des directions admissibles reste malheureusement une question ouverte.

4. Construction de l'approximation initiale. La méthode des directions admissibles ne s'applique que si l'on connaît la valeur de l'approximation initiale dans le domaine D. On obtient celle-ci par les directions admissibles pour le problème de minimisation du nombre η sous les contraintes

$$f_i(x) - \eta \leqslant 0, \ i = 1, \ldots, m, \ Ax - b = 0. \qquad (2.22)$$

Vu qu'il existe un point $\bar{x}$ tel que

$$f_i(\bar{x}) < 0, \ i = 1, \ldots, m, \ A\bar{x} - b = 0,$$

dans les contraintes ci-dessus la valeur minimale de η est strictement inférieure à 0 et après un nombre fini de pas on trouve un point x et un nombre η tels que $\eta < 0$ et les inégalités (2.22) soient vérifiées. Cela signifie que le point x ainsi obtenu satisfait aux contraintes du problème originel et qu'on peut le prendre pour point de départ dans la méthode des directions admissibles.

§ 3. Méthode du gradient conditionnel et méthode de Newton

La méthode du gradient conditionnel est utilisée pour minimiser une fonction non linéaire dans un domaine où le minimum d'une fonction linéaire s'obtient sans trop de peine.

Soit $f(x)$, $x \in E^n$, une fonction continûment dérivable dans un domaine convexe compact Ω. Nous supposerons que le gradient $f'(x)$ de $f(x)$ satisfait dans Ω à la condition de Lipschitz, i.e.

$$\| f'(x_1) - f'(x_2) \| \leqslant L \| x_1 - x_2 \| \tag{3.1}$$

pour tous les points de Ω.

Décrivons la méthode proposée que nous appellerons méthode du gradient conditionnel.

Connaissant x_k, l'approximation au k-ième pas du processus itératif, on calcule $f'(x_k)$ et on trouve le point de minimum de la fonction linéaire $(f'(x_k), z)$ dans Ω. Soit $z(x_k)$ ce point. Posons $p_k = z(x_k) - x_k$ et $x_{k+1} = x_k + \alpha_k p_k$, où $\alpha_k > 0$ est la longueur du pas dans la direction p_k. On recommence avec le point x_{k+1} au départ.

Nous montrerons plus bas qu'en calculant α_k par une règle déterminée il y a convergence, et nous évaluerons la vitesse de convergence. On traitera les mêmes questions pour la méthode de Newton dont la différence de la technique décrite est que la fonction à minimiser est approchée à chaque itération par une forme quadratique (l'approximation est linéaire dans la méthode du gradient conditionnel).

1. Règle régissant le choix de la longueur du pas. Soit x un point quelconque de Ω et soit $z(x)$ un point de minimum de $(f'(x), z)$ dans Ω tel que

$$(f'(x), z(x)) \leqslant (f'(x), z), \quad z \in \Omega. \tag{3.2}$$

Posons $p(x) = z(x) - x$,

$$\eta(x) = \min_{z \in \Omega} (f'(x), z - x) = (f'(x), p(x)).$$

En vertu de (3.2), $\eta(x) \leqslant 0$. Nous voulons évaluer l'accroissement de la fonction lorsqu'on se déplace du point x dans la direction $p(x)$. Utilisons la formule de Taylor et (3.1), il vient

$$f(x + \alpha p(x)) = f(x) + \alpha (f'(\theta), \quad p(x)) = f(x) +$$
$$+ \alpha (f'(x), \quad p(x)) + \alpha (f'(\theta) - f'(x), \quad p(x)) \leqslant f(x) +$$
$$+ \alpha \eta(x) + \alpha^2 L \| p(x) \|^2, \quad \text{où} \quad \theta = x + \xi \alpha p(x), \quad 0 \leqslant \xi \leqslant 1.$$

Ainsi,

$$f(x+\alpha p(x)) \leqslant f(x) + \alpha(\eta(x) + \alpha L \| p(x) \|^2). \quad (3.3)$$

Il en résulte de suite que pour

$$\alpha \leqslant -\frac{1}{2}\frac{\eta(x)}{L\| p(x)\|^2} \quad (3.4)$$

on a l'estimation

$$f(x+\alpha p(x)) \leqslant f(x) + \frac{\alpha\eta(x)}{2}. \quad (3.5)$$

2. Description de l'algorithme. L'algorithme débute avec un point arbitraire x_0 du domaine Ω. Décrivons le pas général.

Supposons x_k déjà construit, $k \geqslant 0$. Après avoir résolu le problème de minimisation de $(f'(x_k), z)$ dans Ω calculons $z(x_k)$, $p(x_k)$, $\eta(x_k)$. Construisons le point $x_{k+1} = x_k + \alpha_k p(x_k)$, avec α_k supposé égal à 2^{-i_0}, i_0 étant le premier indice $i = 0, 1, \ldots$, pour lequel on a l'inégalité

$$f(x_k + 2^{-i}p(x_k)) \leqslant f(x_k) + 2^{-i}\frac{\eta(x_k)}{2}. \quad (3.6)$$

La condition d'arrêt est la suivante: le processus s'arrête sitôt que $\eta(x_k) = 0$.

3. Justification et évaluation de la convergence de l'algorithme. En vertu de la règle gouvernant le choix du pas que nous venons de citer, on a l'inégalité

$$f(x_{k+1}) \leqslant f(x_k) + \frac{\alpha_k\eta(x_k)}{2}. \quad (3.7)$$

Pour justifier la convergence il faut montrer en premier lieu que les inégalités (3.6), (3.7) ont toujours lieu. En effet, en vertu des formules (3.4) et (3.5), l'inégalité (3.6) est vérifiée dès qu'on a

$$2^{-i} \leqslant -\frac{1}{2}\frac{\eta(x_k)}{L\| p(x_k)\|^2},$$

et comme i_0 est le premier indice pour lequel (3.6) est juste,

$$2\alpha_k = 2^{-(i_0-1)} > -\frac{1}{2}\frac{\eta(x_k)}{L\| p(x_k)\|^2},$$

d'où

$$\alpha_k > \frac{1}{4L}\frac{-\eta(x_k)}{\| p(x_k)\|^2}. \quad (3.8)$$

Cela entraîne que si $\eta(x_k) < 0$, l'inégalité (3.6) est vérifiée au bout d'essais en nombre fini et α_k choisi satisfait à (3.8).

LEMME 3.1. *Si x_k, $k = 0, 1, \ldots$, est une suite de points générée dans l'algorithme de gradient conditionnel, alors $x_k \in \Omega$, $f(x_k)$ est monotone décroissante et $\eta(x_k) \to 0$ pour $k \to +\infty$.*

DÉMONSTRATION. Soit $x_k \in \Omega$ pour $k \leqslant m$. Montrons que $x_{m+1} \in \Omega$. En effet, $0 \leqslant \alpha_k \leqslant 1$ et $z(x_k) \in \Omega$. Aussi

$$x_{m+1} = x_m + \alpha_m p(x_m) = x_m + \alpha_m (z(x_m) - x_m) = \\ = (1 - \alpha_m) x_m + \alpha_m z(x_m) \in \Omega,$$

le domaine Ω étant convexe.

Notons que $\| p(x_k) \|$ est borné par une constante C parce que $p(x_k) = z(x_k) - x_k$, $z(x_k) \in \Omega$, $x_k \in \Omega$, et Ω est un compact. Utilisons les formules (3.8), (3.7), il vient

$$f(x_{k+1}) - f(x_k) \leqslant -\frac{1}{8LC^2} \eta^2(x_k). \tag{3.9}$$

La sommation par rapport à tous les $k = 0, 1, \ldots, m-1$ donne

$$f(x_m) - f(x_0) \leqslant -\frac{1}{8LC^2} \sum_{k=0}^{m-1} \eta^2(x_k).$$

Comme le domaine Ω est compact et $f(x)$ continue, on a $f(x_m) \geqslant f_*$ avec f_* la valeur minimale de $f(x)$ dans Ω. Donc

$$\sum_{k=0}^{m-1} \eta^2(x_k) \leqslant 8LC^2 [f(x_0) - f(x_m)] \leqslant 8LC^2 [f(x_0) - f_*].$$

La dernière inégalité entraîne la convergence de la série

$$\sum_{k=0}^{\infty} \eta^2(x_k),$$

ce qui n'est possible que quand $\eta(x_k) \to 0$. Le lemme est démontré.

Les conditions d'arrêt de l'algorithme et les hypothèses du lemme impliquent que dans le cas général ou bien l'algorithme s'arrête après un nombre fini de pas et l'on a la condition $\eta(x_k) = 0$, ou bien on obtient une suite monotone décroissante de valeurs $f(x_k)$ de la fonction $f(x)$.

Dans le premier cas, $\eta(x_k) = 0$ équivaut, en vertu de (3.2), à

$$(f'(x_k), x_k) = (f'(x_k), z(x_k)) \leqslant (f'(x_k), z), \quad z \in \Omega.$$

La dernière relation n'est autre qu'une condition nécessaire de minimum de $f(x)$ en x_k (voir ch. I, § 3).

Le second cas fait l'objet du

LEMME 3.2. *En tout point limite de la suite x_k, $k = 0, 1, \ldots$, on a les conditions nécessaires de minimum de $f(x)$ sur l'ensemble Ω.*

Soit x_* un point limite de la suite x_k, i.e. il existe une suite partielle x_{k_j}, $j \to \infty$, telle que $x_{k_j} \to x_*$. On a les relations suivantes:

$$\eta(x_{k_j}) = (f'(x_{k_j}),\quad z(x_{k_j}) - x_{k_j}),$$
$$(f'(x_{k_j}),\quad z(x_{k_j})) \leqslant (f'(x_{k_j}),\ z),\quad z \in \Omega.$$

On estime sans restreindre la généralité que $z(x_{k_j}) \to z_*$. Comme $\eta(x_k) \to 0$ et $f'(x)$ est une fonction continue de x, ces relations entraînent

$$(f'(x_*),\ z_* - x_*) = 0,$$
$$(f'(x_*),\ z_*) \leqslant (f'(x_*),\ z),\quad z \in \Omega.$$

D'où

$$(f'(x_*),\ x_*) \leqslant (f'(x_*),\ z),\quad z \in \Omega,$$

ce qui démontre le lemme.

THEOREME 3.1. *Etant donné une fonction $f(x)$ convexe, on a*

$$\lim_{k\to\infty} f(x_k) = f_*,$$

où $f_* = \min\limits_{x\in\Omega} f(x)$, et l'estimation

$$f(x_k) - f_* \leqslant \frac{C}{k},$$

avec C une constante positive.

DEMONSTRATION. La convexité de $f(x)$ détermine la validité de l'inégalité

$$f_* - f(x) = f(x_*) - f(x) \geqslant (f'(x),\ x_* - x) \geqslant$$
$$\geqslant \min_{z\in\Omega} (f'(x),\ z - x) = \eta(x).$$

Ainsi, $0 \leqslant f(x) - f_* \leqslant -\eta(x)$. Quel que soit k, on a donc

$$0 \leqslant f(x_k) - f_* \leqslant -\eta(x_k). \tag{3.10}$$

Le lemme 3.1 entraîne $\eta(x_k) \to 0$. Aussi la dernière inégalité montre que $f(x_k) \to f_*$, ce qui démontre la première partie du théorème.

Réunissons (3.9) et (3.10), il vient

$$(f(x_{k+1}) - f_*) - (f(x_k) - f_*) \leqslant -\frac{1}{8LC^2}[f(x_k) - f_*]^2.$$

En notant $\varphi_k = f(x_k) - f_*$ nous obtenons

$$\varphi_{k+1} \leqslant \varphi_k\left(1 - \frac{\varphi_k}{8LC^2}\right)$$

ou

$$\varphi_{k+1} \leqslant \varphi_k(1 - \varkappa\varphi_k),\quad \varkappa = \frac{1}{8LC^2}.$$

Posons maintenant $\varphi_k = \frac{\gamma_k}{k}$, nous avons

$$\frac{\gamma_{k+1}}{\gamma_k} \leqslant \frac{k+1}{k}\left(1 - \varkappa \frac{\gamma_k}{k}\right) \tag{3.11}$$

ou

$$\frac{\gamma_{k+1}}{\gamma_k} \leqslant 1 + \frac{1}{k} - \frac{\varkappa (k+1)\gamma_k}{k^2}.$$

Pour tout k deux cas se présentent.

1) $\frac{\gamma_{k+1}}{\gamma_k} \leqslant 1$, i.e. $\gamma_{k+1} \leqslant \gamma_k$.

2) $\frac{\gamma_{k+1}}{\gamma_k} > 1$.

Alors $\frac{1}{k} - \varkappa \frac{k+1}{k^2}\gamma_k > 0$, i.e.

$$\gamma_k < \frac{1}{\varkappa}\frac{k}{k+1} < \frac{1}{\varkappa}.$$

On obtient à partir de (3.11)

$$\frac{\gamma_{k+1}}{\gamma_k} \leqslant \frac{k+1}{k} \leqslant 2$$

pour $k \geqslant 1$. On ne peut se placer alors que dans deux situations suivantes.

1) Il existe un nombre fini d'indices k tels que $\gamma_k < \frac{1}{\varkappa}$. En vertu de ce que nous avons dit, la suite γ_k ne croît de façon monotone pour aucun k grand, i.e. elle reste bornée.

2) Il existe une infinité d'indices k tels que $\gamma_k < \frac{1}{\varkappa}$. Nous désignerons par $\mathcal{Y}$ l'ensemble de tels k de sorte que $\gamma_k \leqslant \frac{1}{\varkappa}$ pour $k \in \mathcal{Y}$. Soit $j \notin \mathcal{Y}$. On peut trouver alors deux indices $k_1, k_2 \in \mathcal{Y}$, tels que $k_1 < j < k_2$ et $k \notin \mathcal{Y}$ pour tous les $k_1 < k < k_2$. Alors

$$\gamma_{k_1+1} \leqslant 2\gamma_{k_1} \leqslant \frac{2}{\varkappa}$$

et $\gamma_{i+1} \leqslant \gamma_i$ quels que soient $i = k_1 + 1, \ldots, k_2 - 1$. Donc

$$\gamma_j \leqslant \frac{2}{\varkappa}, \quad j \notin \mathcal{Y}.$$

Nous avons donc montré que la suite est également bornée par une constante C. Il en résulte

$$\varphi_k = \frac{\gamma_k}{k} \leqslant \frac{C}{k},$$

ce qui achève la démonstration du théorème.

Cette estimation révèle une convergence pas trop rapide de l'algorithme. Il est vrai qu'il s'agit de la majoration, ce qui fait espérer une vitesse plus grande. Or, cela ne se passe malheureusement pas ainsi dans le cas général. Selon Cannon et Cullum [47] l'estimation obtenue est précise s'il s'agit de minimiser une fonction convexe sur un polyèdre.

4. Evaluation de la convergence dans le cas d'un domaine fortement convexe. Soit Ω un domaine fortement convexe, i.e. il existe un nombre $\delta > 0$ tel que, quels que soient $x, y \in \Omega$, les points $\frac{x+y}{2} + w$ appartiennent à Ω pour tous les w vérifiant $\| w \| \leqslant \delta \| x - y \|^2$. Alors

$$\eta(x) = \min_{z \in \Omega} (f'(x), z - x) \leqslant \min_{\|w\| \leqslant \delta \|z(x) - x\|^2} \left(f'(x), \frac{z(x) + x}{2} + w - x \right) \leqslant$$

$$\leqslant \frac{1}{2} (f'(x), z(x) - x) - \delta \| z(x) - x \|^2 \| f'(x) \|.$$

D'où

$$\frac{1}{2} \eta(x) \leqslant -\delta \| f'(x) \| \| z(x) - x \|^2,$$

ou

$$-\frac{1}{2} \frac{\eta(x)}{\| p(x) \|^2} \geqslant \delta \| f'(x) \|. \tag{3.12}$$

THÉORÈME 3.2. *Si $f(x)$ est une fonction convexe, Ω un domaine fortement convexe et $\| f'(x) \| \geqslant \varepsilon_0 > 0$ quels que soient $x \in \Omega$, la méthode du gradient conditionnel converge en progression géométrique, i.e. $\| x_k - x_* \| \leqslant C q_0^k$, $q_0 < 1$.*

DÉMONSTRATION. On obtient à partir de (3.7) et (3.8)

$$\varphi_k - \varphi_{k+1} = f(x_k) - f(x_{k+1}) \geqslant \frac{1}{8L} \frac{\eta^2(x_k)}{\| p(x_k) \|^2}.$$

Utilisons (3.12) et (3.10), il vient

$$\varphi_k - \varphi_{k+1} \geqslant \frac{1}{4L} \delta \varepsilon_0 (-\eta(x)) \geqslant \frac{\delta \varepsilon_0}{4L} \varphi_k,$$

c'est-à-dire

$$\varphi_{k+1} \leqslant \left(1 - \frac{\delta \varepsilon_0}{4L} \right) \varphi_k.$$

Donc

$$\varphi_m \leqslant q^m \varphi_0, \quad q = 1 - \frac{\delta \varepsilon_0}{4L} < 1.$$

En vertu des conditions nécessaires et suffisantes de minimum,

$$(f'(x_*), x - x_*) \geqslant 0.$$

On a donc pour tous les w, $\| w \| \leqslant \delta \| x - x_* \|^2$,

$$\left(f'(x_*),\ \frac{x+x_*}{2} + w - x_*\right) \geqslant 0,$$

d'où

$$\frac{1}{2}(f'(x_*), x - x_*) \geqslant \delta \| x - x_* \|^2 \| f'(x_*) \|.$$

Mais $\| f'(x_*) \| \geqslant \varepsilon_0$ et, par suite de la convexité de $f(x)$,

$$f(x) - f(x_*) \geqslant (f'(x_*), x - x_*).$$

On a finalement

$$f(x) - f(x_*) \geqslant 2\delta\varepsilon_0 \| x - x_* \|^2. \tag{3.13}$$

D'où

$$\| x_k - x_* \| \leqslant \left(\frac{1}{2\delta\varepsilon_0}\right)^{1/2} \varphi_k^{1/2} \leqslant \left(\frac{\varphi_0}{2\delta\varepsilon_0}\right)^{1/2} (q^{1/2})^k.$$

Notons

$$C = \left(\frac{\varphi_0}{2\delta\varepsilon_0}\right)^{1/2}, \quad q_0 = \left(1 - \frac{\delta\varepsilon_0}{4L}\right)^{1/2},$$

il vient

$$\| x_k - x_* \| \leqslant C q_0^k,$$

c.q.f.d.

5. **Méthode de Newton à pas variable.** Considérons le problème de minimisation d'une fonction convexe régulière $f(x)$ sur un ensemble (convexe, compact) Ω.

On le résout moyennant le processus itératif

$$x_{k+1} = x_k + \alpha_k p_k, \ \alpha_k > 0, \tag{3.14}$$

où la direction de déplacement $p_k = \bar{x}_k - x_k$ est solution du problème de minimisation sur Ω de la fonction quadratique

$$\psi_k(x) = (f'(x_k),\ x - x_k) + \frac{1}{2}(f''(x_k)(x - x_k),\ x - x_k),$$

et où l'on prend pour α_k la plus grande valeur du paramètre α obtenue par fractionnement à partir de $\alpha = 1$ et vérifiant l'inégalité

$$f(x_k + \alpha p_k) - f(x_k) \leqslant \varepsilon\alpha\psi_k(\bar{x}_k), \quad 0 < \varepsilon < 1. \tag{3.15}$$

Le choix de α_k peut s'effectuer d'une autre manière (comme on l'a décrit au § 2, ch. II, procédés (II.2.2), (II.2.3)).

Nous verrons que dans des conditions déterminées la convergence de la méthode de Newton est soit superlinéaire soit quadratique. Dans le cas où la recherche du minimum de $\psi_k(x)$, $x \in \Omega$, ne présente pas de difficulté, l'efficacité de la technique proposée s'avère donc considérable.

6. Propriétés de la méthode de Newton.

THÉORÈME 3.3. *Si l'on recherche le minimum d'une fonction convexe deux fois continûment dérivable $f(x)$ sur un ensemble convexe fermé borné Ω par la méthode* (3.14), *où α_k et p_k se définissent comme décrit ci-dessus, alors (pour tout choix de l'approximation initiale $x_0 \in \Omega$):*

1) *$f(x_k)$ est monotone décroissante;*
2) $\lim\limits_{k\to\infty} f(x_k) = f(x_*) = \min\limits_{x\in\Omega} f(x)$.

DÉMONSTRATION. Une fonction continue $\psi_k(x)$ admet bien un point de minimum $\bar{x}_k$ (peut-être non unique) sur le compact Ω (théorème de Weierstrass). Quel que soit k, le point x_{k+1} appartient à Ω parce que $x_{k+1} = x_k + \alpha_k(\bar{x}_k - x_k) = \alpha_k\bar{x}_k + (1-\alpha_k)x_k$ et $\alpha_k \in [0, 1]$. Etant donné la convexité de $\psi_k(x)$, on a $\psi_k(x_{k+1}) = \psi_k(\alpha_k\bar{x}_k + (1-\alpha_k)x_k) \leqslant \alpha_k\psi_k(\bar{x}_k) + (1-\alpha_k)\psi_k(x_k)$. Mais $\psi_k(x_k) = 0$, donc

$$\psi_k(x_{k+1}) \leqslant \alpha_k\psi_k(\bar{x}_k). \tag{3.16}$$

Utilisons la formule de Taylor et (3.16), il vient

$$f(x_{k+1}) - f(x_k) = \psi_k(x_{k+1}) + \frac{\alpha_k^2}{2}(F_k p_k, p_k) \leqslant$$
$$\leqslant \alpha_k\psi_k(\bar{x}_k)\left(1 + \frac{\alpha_k}{2}\frac{\|F_k\|\|p_k\|^2}{\psi_k(\bar{x}_k)}\right), \tag{3.17}$$

avec $F_k = f''(x_{kc}) - f''(x_k)$, $x_{kc} = x_k + \theta(x_{k+1} - x_k)$, $\theta \in [0, 1]$. Il s'ensuit que si $\psi_k(\bar{x}_k) \neq 0$ (auquel cas $\psi_k(\bar{x}_k) < \psi_k(x_k) = 0$), on a pour un certain $\alpha_k > 0$

$$1 + \frac{\alpha_k}{2}\frac{\|F_k\|\|p_k\|^2}{\psi_k(\bar{x}_k)} \geqslant \varepsilon \tag{3.18}$$

et, partant, (3.15), ce qui démontre la possibilité de choisir α_k de façon décrite.

L'inégalité (3.15) entraîne $f(x_{k+1}) \leqslant f(x_k)$. Montrons que $\psi_k(\bar{x}_k)$ tend vers zéro pour $k \to \infty$. La fonction continue $f''(x)$ est bornée sur Ω fermé borné: $\|f''(x)\| \leqslant M$. Par conséquent, $\|F_k\| \leqslant 2M$. Le vecteur p_k est borné lui aussi: $\|p_k\| \leqslant \max\limits_{x,y\in\Omega}\|x - y\| = d$. Supposons que, quel que soit k, $\psi_k(\bar{x}_k) \leqslant -\beta < 0$. Alors

$$1 + \frac{\alpha_k}{2}\frac{\|F_k\|\|p_k\|^2}{\psi_k(\bar{x}_k)} \geqslant 1 - \alpha_k\frac{Md^2}{\beta},$$

et l'inégalité (3.18) (et donc (3.15)) a manifestement lieu dès que $\alpha_k = \frac{(1-\varepsilon)\beta}{Md^2} = C > 0$. Ceci étant, il découle de (3.15) que

$f(x_{k+1}) - f(x_k) \leqslant -\varepsilon C\beta$ pour tout k, ce qui contredit la propriété de la fonction $f(x)$ d'être minorée sur le compact Ω.

Ainsi, on n'a $\psi_k(\bar{x}_k) \leqslant -\beta$ que pour certains k, i.e. $\psi_k(\bar{x}_k) \to 0$ au moins pour $k \to \infty$. Cela signifie que tout point limite de la suite (3.14) vérifie une condition nécessaire (et suffisante, $f(x)$ étant convexe) de minimum de $f(x)$ sur l'ensemble Ω (voir ch. I, § 4). Si l'on en tient compte, la démonstration de la dernière affirmation du théorème s'effectue comme dans le théorème 3.1.

Le théorème 3.3 montre qu'à la différence des problèmes de minimisation sans contraintes où la méthode de Newton ne joue que pour des fonctions fortement convexes, on peut l'appliquer, grâce au caractère borné de Ω, à une fonction convexe. Il n'est pas moins vrai que cette méthode est particulièrement indiquée quand il s'agit de minimiser des fonctions fortement convexes, car c'est alors qu'elle a une convergence très rapide.

THÉORÈME 3.4. *Si, dans les conditions du théorème* 3.3, $f(x)$ *est fortement convexe, i.e.*

$$m \| y \|^2 \leqslant (f''(x)\, y, y) \leqslant M \| y \|^2, \quad m > 0, \quad x \in \Omega, \quad y \in E^n, \tag{3.19}$$

la suite (3.14) *converge vers la solution et la convergence est superlinéaire* (*autrement dit, on a l'estimation* (II.2.5)).

DÉMONSTRATION. L'existence et l'unicité de la solution du problème considéré découlent des résultats généraux du § 3, ch. I. Le point $\bar{x}_k$ vérifie la condition nécessaire de minimum de $\psi_k(x)$ sur Ω (§ 4, ch. I):

$$(\psi_k'(\bar{x}_k), \bar{x}_k - x_k) \leqslant 0,$$

i.e.

$$(f'(x_k), \bar{x}_k - x_k) + (f''(x_k)(\bar{x}_k - x_k), \bar{x}_k - x_k) \leqslant 0,$$

d'où

$$(f'(x_k), p_k) \leqslant -(f''(x_k)\, p_k, p_k). \tag{3.20}$$

Par suite de (3.20) et de la première estimation (3.19),

$$\psi_k(\bar{x}_k) \leqslant -\frac{m}{2} \| p_k \|^2. \tag{3.21}$$

Reprenons (3.17) avec cette estimation:

$$f(x_{k+1}) - f(x_k) \leqslant \alpha_k \psi_k(\bar{x}_k) \left(1 - \frac{\alpha_k \| F_k \|}{m}\right). \tag{3.22}$$

Comme $\psi_k(\bar{x}_k) \to 0$ (théorème 3.3), (3.21) entraîne $\| p_k \| \to 0$ quand $k \to \infty$. Il en résulte, vu la continuité uniforme de la dérivée seconde $f''(x)$ sur Ω, que $\| F_k \| \to 0$. Ceci étant, il découle de (3.22) qu'à partir d'un $k = N_1(\varepsilon)$ l'inégalité (3.15) est satisfaite pour $\alpha_k \equiv 1$, i.e. la méthode (3.14) devient la méthode ordinaire de Newton à pas unité. Quand $k > N_1(\varepsilon)$, on a, compte tenu de la

convexité de $\psi_k(x)$,

$$\psi_k(\bar{x}_k) = \psi_k(x_{k+1}) \geqslant (f'(x_k),\quad x_{k+1} - x_k) =$$
$$= (f'(x_{k-1}),\quad x_{k+1} - x_k) + (f'(x_k) - f'(x_{k-1}),\quad x_{k+1} - x_k).$$

La transformation du dernier produit scalaire par la formule de Lagrange généralisée (ch. I, § 5) et certaines autres transformations donnent

$$\psi_k(\bar{x}_k) = (f'(x_{k-1}) + f''(x_{k-1})(x_k - x_{k-1}),\quad x_{k+1} - x_k) +$$
$$+ (\Phi_k(x_k - x_{k-1}),\quad x_{k+1} - x_k), \tag{3.23}$$

où

$$\Phi_k = f''(x_{k-1} + \theta_1(x_k - x_{k-1})) - f''(x_{k-1}),\quad \theta \quad [0, 1].$$

Notons que $(f'(x_{k-1}) + f''(x_{k-1})(x_k - x_{k-1}),\ x_{k+1} - x_k) = (\psi'_{k-1}(x_k),\ x_{k+1} - x_k)$. Comme $\psi_{k-1}(x_k) = \min\limits_{x \in \Omega} \psi_{k-1}(x)$, on a, quel que soit $x \in \Omega$, $(\psi'_{k-1}(x_k),\ x - x_k) \geqslant 0$ (une condition nécessaire de minimum). On a donc $(\psi'_{k-1}(x_k),\ x_{k+1} - x_k) \geqslant 0$ et (3.23) entraîne

$$-\psi_k(\bar{x}_k) \leqslant \|\Phi_k\| \, \|x_k - x_{k-1}\| \, \|x_k - x_{k+1}\| = \|\Phi_k\| \, p_{k-1} \|p_k\|. \tag{3.24}$$

Comparons les estimations (3.21) et (3.24), il vient

$$\|x_{k+1} - x_k\| \leqslant \frac{2\|\Phi_k\|}{m} \|x_k - x_{k-1}\|. \tag{3.25}$$

$f''(x)$ étant uniformément continue sur Ω, on a $\|\Phi_k\| \to 0$. Il existe donc un nombre $N(\varepsilon)$ tel que $\lambda_k = \dfrac{2\|\Phi_k\|}{m} < 1$ pour $k \geqslant N(\varepsilon)$. Posons $\|x_N - x_{N-1}\| = C_1$, $1 - \lambda_N = \gamma > 0$. Alors

$$\|x_i - x_{N+l}\| \leqslant \sum_{k=N+l}^{i-1} \|x_{k+1} - x_k\| \leqslant$$
$$\leqslant C_1 \lambda_N \lambda_{N+1} \ldots \lambda_{N+l} (1 + \lambda_{N+l+1} + \lambda^2_{N+l+1} + \ldots + \lambda^{i-(N+l+1)}_{N+l+1}) \leqslant$$
$$\leqslant \frac{C_1}{1 - \lambda_N} \lambda_N \ldots \lambda_{N+l} = C \lambda_N \ldots \lambda_{N+l}.$$

Par conséquent, $\|x_i - x_{N+l}\| \to 0$ quand $i, l \to \infty$, i.e. la suite $\{x_k\}$ est fondamentale et, vu la complétude de l'espace E^n, possède une limite $x_* \in \Omega$, et

$$\|x_{N+l} - x_*\| \leqslant C \lambda_N \lambda_{N+1} \ldots \lambda_{N+l}. \tag{3.26}$$

En vertu du théorème 3.3, $\lim\limits_{k \to \infty} f(x_k) = f(x_*) = \min\limits_{x \in \Omega} f(x)$.

Ainsi, la suite (3.14) converge vers la solution et, selon l'estimation (3.26), cette convergence est superlinéaire. Le théorème est démontré.

Si la dérivée seconde $f''(x)$ vérifie sur Ω la condition de Lipschitz de constante R, l'inégalité (3.25) prend la forme

$$\| x_{k+1} - x_k \| \leqslant \frac{2R}{m} \| x_k - x_{k-1} \|^2. \tag{3.27}$$

Notons $\beta_k = \frac{2R}{m} \| x_{k+1} - x_k \|$. Comme $\| x_{k+1} - x_k \| \to 0$, il existe un nombre $L(\varepsilon)$ tel que $\beta_k < 1$ pour $k > L(\varepsilon)$. Compte tenu de (3.27) on a, quand $k > L$,

$$\beta_k \leqslant \beta_{k-1}^2 \leqslant \ldots \leqslant \beta_L^{2^{k-L}}.$$

Donc, quel que soit $i > L + l$, $l = 0, 1, \ldots$,

$$\| x_i - x_{L+l} \| \leqslant \sum_{k=L+l}^{i-1} \| x_{k+1} - x_k \| \leqslant \frac{m}{2R} \sum_{k=L+l}^{i-1} \beta_L^{2^{k-L}} = \frac{m}{2R} \sum_{s=l}^{i-1} \beta_L^{2^s}.$$

Puisque $x_k \to x_*$, on a $\| x_{L+l} - x_* \| = \lim_{i \to \infty} \| x_{L+l} - x_i \|$, i.e.

$$\| x_{L+l} - x_* \| \leqslant \frac{m}{2R} \sum_{s=l}^{\infty} \beta_L^{2^s},$$

estimation qui s'écrit encore (étant donné la convergence de la série $\sum_{s=l}^{\infty} \beta_L^{2^s}$

$$\| x_{L+l} - x_* \| \leqslant C \beta_L^{2^l}, \quad C < \infty.$$

Cette estimation signifie qu'on a le

THÉORÈME 3.5. *Si, dans les conditions du théorème* 3.4, *la matrice* $f''(x)$ *vérifie sur l'ensemble* Ω *la condition de Lipschitz de constante* R, *la convergence de la suite* (3.14) (*avec* α_k *et* p_k *choisis de façon décrite ci-dessus*) *vers la solution est quadratique.*

Voyons maintenant les propriétés de la méthode de Newton lorsque α_k est choisi à partir de la condition de minimum de $f(x)$ dans la direction de déplacement.

On ne saurait dans ce cas raisonner comme sur l'évaluation de la convergence dans les problèmes sans contraintes (§ 2, ch. II) parce que la seconde estimation (II.1.11) n'est plus valable.

LEMME 3.3. *Si on a à minimiser une fonction* $f(x)$ *vérifiant les conditions* (3.19) *et si, dans la méthode* (3.14), α_k *est choisi à partir de la condition*

$$f(x_k + \alpha_k p_k) = \min_{0 \leqslant \alpha \leqslant 1} f(x_k + \alpha p_k), \tag{3.28}$$

alors $x_k \to x_*$ *et* $\alpha_k \to 1$ *pour* $k \to \infty$.

DÉMONSTRATION. D'après la formule de Taylor,

$$f(x_{k+1})-f(x_k)=\alpha_k(f'(x_k),\ p_k)+\frac{\alpha_k^2}{2}(f''(x_{kc})\,p_k,\ p_k).$$

Pour α_k vérifiant (3.28) le second membre considéré comme fonction de la variable α doit atteindre son minimum. Si l'on prend en considération les estimations (3.19) et (3.20), on se convainc donc sans peine que

$$1\geqslant\alpha_k\geqslant-\frac{(f'(x_k),\ p_k)}{M\,\|p_k\|^2}\geqslant\frac{m\|p_k\|^2}{M\,\|p_k\|^2}=\frac{m}{M}.$$

Ainsi, $\alpha_k\geqslant C>0$ et on démontre comme pour le théorème 3.3 que $\psi_k(\bar{x}_k)\to 0$, i.e. la suite (3.14) avec α_k choisi à partir de (3.28) converge vers la solution. Ceci étant, $\|p_k\|\to 0$ et $\|F_k\|\to 0$ (théorème 3.4).

Montrons que $\alpha_k\to 1$:

$$f(x_{k+1})=\psi_k(x_{k+1})+\frac{\alpha_k^2}{2}(F_kp_k,\ p_k)=\psi_k(\bar{x}_k)+$$

$$+(\psi_k'(\bar{x}_k),\ x_{k+1}-\bar{x}_k)+\frac{1}{2}(\psi_k''(\bar{x}_k)(x_{k+1}-\bar{x}_k),\ x_{k+1}-\bar{x}_k)+$$

$$+\frac{\alpha_k^2}{2}(F_kp_k,\ p_k).$$

Etant donné que $x_{k+1}-\bar{x}_k=(\alpha_k-1)\,p_k$, on a

$$f(x_{k+1})=\psi_k(\bar{x}_k)+(\psi_k'(\bar{x}_k),\ x_{k+1}-\bar{x}_k)+$$

$$+\frac{(\alpha_k-1)^2}{2}(\psi_k''(\bar{x}_k)\,p_k,\ p_k)+\frac{\alpha_k^2}{2}(F_kp_k,\ p_k).$$

Notons que $(\psi_k'(\bar{x}_k),\ x_{k+1}-\bar{x}_k)\geqslant 0$ et $(\psi_k''(\bar{x}_k)\,p_k,\ p_k)=$ $=(f''(x_k)\,p_k,\ p_k)\geqslant m\,\|p_k\|^2$. Par ailleurs, $(F_kp_k,\ p_k)=o(\|p_k\|^2)$ (du moment que $\|F_k\|\to 0$ quand $\|p_k\|\to 0$). Le minimum de la différence $f(x_{k+1})-\psi_k(\bar{x}_k)$ n'est donc atteint que lorsque $\alpha_k\to 1$; dans le cas contraire, on aurait $1-\alpha_k\geqslant\beta>0$ pour tout k et $f(x_{k+1})-\psi_k(\bar{x}_k)=O(\|p_k\|^2)>0$ tandis que, lorsque $\alpha=1$, on a $f(\bar{x}_k)-\psi_k(\bar{x}_k)=\frac{1}{2}(F_kp_k,\ p_k)=o(\|p_k\|^2)$, i.e. pour un k suffisamment grand on aurait nécessairement $f(\bar{x}_k)<f(x_{k+1})$, ce qui contredit la condition de choix de α_k. Le lemme est démontré.

THÉORÈME 3.6. *Si $f(x)$ est une fonction vérifiant les exigences du théorème* 3.4 *et le paramètre α_k est choisi dans la méthode* (3.14) *à partir de la condition* (3.28), *alors $x_k\to x_*$ est superlinéaire.*

DÉMONSTRATION. En vertu des estimations (3.16) et (3.21)

$$\psi_k(x_{k+1}) \leqslant \alpha_k \psi_k(\bar{x}_k) \leqslant \alpha_k^2 \psi_k(\bar{x}_k) \leqslant -\frac{\alpha_k^2}{2} m \| p_k \|^2,$$

i.e.

$$\psi_k(x_{k+1}) \leqslant -\frac{m}{2} \| x_{k+1} - x_k \|^2. \tag{3.29}$$

D'autre part $\psi_k(x_{k+1}) \geqslant (f'(x_k), x_{k+1} - x_k) = (f'(x_k), x_{k+1} - \bar{x}_{k-1}) + (f'(x_k), \bar{x}_{k-1} - x_k)$. Puisque x_k réalise le minimum de $f(x)$ dans la direction $\bar{x}_{k-1} - x_{k-1}$, on a $(f'(x_k), \bar{x}_{k-1} - x_k) \geqslant 0$. Par suite,

$$\psi_k(x_{k+1}) \geqslant (f'(x_k), x_{k+1} - \bar{x}_{k-1}) =$$
$$= (f'(x_{k-1}), x_{k+1} - \bar{x}_{k-1}) + (f'(x_k) - f'(x_{k-1}), x_{k+1} - \bar{x}_{k-1}).$$

Exprimant le dernier terme du second membre moyennant la formule de Lagrange pour les opérateurs, on obtient par certaines transformations

$$\psi_k(x_{k+1}) \geqslant (f'(x_{k-1}) + f''(x_{k-1})(x_k - x_{k-1}), x_{k+1} - \bar{x}_{k-1}) +$$
$$+ (\Phi_k(x_k - x_{k-1}), x_{k+1} - \bar{x}_{k-1}),$$

où $\Phi_k = f''(x_{k-1} + \theta(x_k - x_{k-1})) - f''(x_{k-1})$, $\theta \in [0, 1]$. Vu que $x_k - x_{k-1} = \alpha_{k-1}(\bar{x}_{k-1} - x_{k-1})$, il se trouve que

$$\psi_k(x_{k+1}) \geqslant (f'(x_{k-1}) + f''(x_{k-1})(\bar{x}_{k-1} - x_{k-1}), x_{k+1} - \bar{x}_{k-1}) +$$
$$+ ((\alpha_{k-1} - 1) f''(x_{k-1})(\bar{x}_{k-1} - x_{k-1}) + \Phi_k(x_k - x_{k-1}), x_{k+1} - \bar{x}_{k-1}).$$

Comme

$$\psi_{k-1}(\bar{x}_{k-1}) = \min_{x \in \Omega} \psi_{k-1}(x),$$

on a $(\psi'_{k-1}(\bar{x}_{k-1}), \bar{x}_{k-1} - x) \leqslant 0$ quels que soient $x \in \Omega$ (une condition nécessaire de minimum). Par conséquent

$$(\psi'_{k-1}(\bar{x}_{k-1}), \bar{x}_{k-1} - x_{k+1}) =$$
$$= (f'(x_{k-1}) + f''(x_{k-1})(\bar{x}_{k-1} - x_{k-1}), \bar{x}_{k-1} - x_{k+1}) \leqslant 0.$$

Compte tenu de cette inégalité établissons l'estimation

$$\psi_k(x_{k+1}) \geqslant \left(\left(\frac{\alpha_{k-1} - 1}{\alpha_{k-1}} f''(x_{k-1}) + \Phi_k \right)(x_k - x_{k-1}), x_{k+1} - \bar{x}_{k-1} \right).$$

D'où, en notant $\| [(\alpha_{k-1} - 1)/\alpha_{k-1}] f''(x_{k-1}) + \Phi_k \| = b_k$,

$$-\psi_k(x_{k+1}) \leqslant b_k \| x_k - x_{k-1} \| \| x_{k+1} - \bar{x}_{k-1} \| \leqslant$$
$$\leqslant b_k \| x_k - x_{k-1} \| (\| x_{k+1} - x_k \| + \| \bar{x}_{k-1} - x_k \|).$$

Rappelons que

$$\bar{x}_{k-1} - x_k = \frac{1-\alpha_{k-1}}{\alpha_{k-1}} (x_k - x_{k-1})$$

et notons $[(1-\alpha_{k-1})/\alpha_{k-1}]\, b_k = c_k$, il vient

$$-\psi_k(x_{k+1}) \leqslant b_k \| x_k - x_{k-1} \| \, \| x_{k+1} - x_k \| + c_k \| x_k - x_{k-1} \|^2.$$

Puisque $\alpha_k \to 1$, $\| p_k \| \to 0$ (lemme 3.3), on a $b_k \to 0$, $c_k \to 0$. Comparons l'estimation obtenue et (3.29), nous avons

$$\| x_{k+1} - x_k \|^2 \leqslant \xi_k \| x_k - x_{k-1} \| \, \| x_{k+1} - x_k \| + \rho_k \| x_k - x_{k-1} \|^2,$$

où

$$\xi_k = \frac{2b_k}{m}, \quad \rho_k = \frac{2c_k}{m}.$$

Résolvons enfin l'inégalité quadratique obtenue par rapport à $\| x_{k+1} - x_k \|$:

$$\| x_{k+1} - x_k \| \leqslant \mu_k \| x_k - x_{k-1} \|,$$

avec

$$\mu_k = \frac{\xi_k}{2} + \sqrt{\frac{\xi_k^2}{4} + \rho_k} \to 0 \quad \text{pour } k \to \infty.$$

La démonstration ultérieure imite celle du théorème 3.4.

§ 4. Méthode du plan sécant

La méthode du plan sécant est destinée à résoudre les problèmes de programmation convexe. Son idée fondamentale est la suivante: le domaine admissible est approché par un polyèdre qui diminue d'un pas d'itération à l'autre et approche toujours mieux le domaine admissible dans le voisinage de la solution.

La méthode s'applique au problème de programmation convexe de la forme: minimiser $f_0(x) = (c, x)$ avec la contrainte

$$f(x) \leqslant 0, \tag{4.1}$$

où $f(x)$ est une fonction convexe continue.

Le fait que la fonction à minimiser $f_0(x)$ est linéaire et que la contrainte (4.1) se réduit à une seule inégalité ne joue pas un grand rôle car si le domaine est donné par plusieurs inégalités

$$f_i(x) \leqslant 0, \; i = 1, \ldots, m, \tag{4.2}$$

$f_i(x)$ étant convexes, ce système se récrit comme (4.1) en posant

$$f(x) = \max_{1 \leqslant i \leqslant m} f_i(x).$$

Dans le cas de $f_0(x)$ convexe mais non linéaire, on introduit une coordonnée supplémentaire x^{n+1} et on ajoute l'inégalité

$$f_{m+1}(x, x^{n+1}) = f_0(x) - x^{n+1} \leqslant 0$$

au système (4.2) pour se ramener à la minimisation de la fonction linéaire x^{n+1} dans les contraintes (4.2). Le problème sera donc abordé sous forme (4.1). Avant de passer à l'algorithme même rappelons que a est un vecteur d'appui à $f(x)$ au point x_0 si $f(x) \geqslant f(x_0) + (a, x - x_0)$ pour tous les x. Les résultats du § 2, ch. I entraînent la non-vacuité de l'ensemble de tels vecteurs en tout point de l'espace pour une fonction convexe continue.

1. **Algorithme.** Soit

$$\Omega = \{x : f(x) \leqslant 0\}$$

un domaine admissible non vide. Supposons de plus que Ω est compact et qu'il existe des vecteurs a_k, $k = -l, -(l-1), \ldots, -1, 0$, et des nombres b_k tels que le domaine

$$S = \{x : (a_k, x) - b_k \leqslant 0,\ k = -l, \ldots, 0\}$$

soit compact et contienne Ω.

Lorsque $k \geqslant 0$, les approximations successives s'obtiennent selon la règle suivante. Posons $S_0 = S$. Si S_k est déjà construit, x_k est une solution quelconque du problème de programmation linéaire : minimiser $f_0(x) = (c, x)$ pour $x \in S_k$. La construction du domaine suivant S_{k+1} obéit à la règle

$$S_{k+1} = \{x : (a_{k+1}, x) - b_{k+1} \leqslant 0\} \cap S_k, \tag{4.3}$$

avec a_{k+1} un vecteur d'appui à $f(x)$ en x_k et

$$b_{k+1} = (a_{k+1}, x_k) - f(x_k). \tag{4.4}$$

Il résulte de (4.3) que $S_{k+1} \subset S_k$ et, quand $k \geqslant 1$,

$$\begin{aligned} S_k = \{x :\ & (a_j, x) - b_j \leqslant 0, \\ & j = -l, \ldots, -1, 0, \ldots, k-1\}. \end{aligned} \tag{4.5}$$

Lemme 4.1. *Quels que soient $k \geqslant 1$, on a $\Omega \subset S_k$.*

Démonstration. Soit $x \in \Omega$, i.e. $f(x) \leqslant 0$. Alors

$$f(x) \geqslant f(x_{j-1}) + (a_j, x - x_{j-1}) = (a_j, x) - b_j$$

et donc $(a_j, x) - b_j \leqslant 0$, $j = 1, \ldots, k$. Quand $j \leqslant 0$, les dernières inégalités sont satisfaites en vertu du choix de a_j et b_j pour $j \leqslant 0$. Le lemme est démontré.

Le lemme 4.1 implique de suite

$$f_0(x_0) \leqslant f_0(x_1) \leqslant \ldots \leqslant f_0(x_k) \leqslant f_0(x_{k+1}) \leqslant \ldots$$

D'autre part, si x_* est un point de minimum de $f_0(x)$ dans Ω, alors $f_0(x_k) \leqslant f_0(x_*)$, car $S_k \supset \Omega$.

THÉORÈME 4.1. *Etant donné une fonction $f(x)$ convexe continue, un domaine Ω compact et un nombre K tel que, quel que soit $x \in S$, le vecteur a qui est, s'il existe vecteur d'appui à $f(x)$ en x vérifie l'inégalité $\| a \| \leqslant K$, tout point limite x_* de la suite x_k, $k = 0, 1, \ldots$, est solution du problème* (4.1) *et $f(x_k) \to 0$.*

DÉMONSTRATION. Puisque $S_0 = S$, $S_k \supset S_{k+1}$, la suite x_k tout entière appartient au compact S et elle possède donc toujours des points limites.

Notons à présent que si $f(x_k) \leqslant 0$ pour un k, alors $x_k \in \Omega$ et, partant, $f_0(x_k) \geqslant f_0(x_*)$. Nous avons cependant montré que $f_0(x_k) \leqslant \leqslant f_0(x_*)$. Ainsi, $f_0(x_k) = f_0(x_*)$, i.e. x_k est solution du problème originel.

Supposons la suite x_k infinie et $f(x_k) > 0$ pour tous les k. Démontrons $f(x_k) \to 0$. Admettons le contraire. Il existe alors un nombre $r > 0$ et une suite partielle d'indices k (notons-la $\mathcal{J}$) tels que $f(x_k) \geqslant \geqslant r$, $k \in \mathcal{J}$. On estime, sans restreindre la généralité, que $x_k \to \bar{x}$, $k \in \mathcal{J}$, du moment que la suite x_k est élément d'un ensemble compact.

Soit maintenant k et $j \in \mathcal{J}$ et $k > j$. Le point x_k vérifie alors par construction l'inégalité

$$(a_{j+1}, x_k) - b_j = (a_{j+1}, x_k - x_j) + f(x_j) \leqslant 0,$$

d'où

$$f(x_j) \leqslant (a_{j+1}, x_j - x_k) \leqslant K \| x_j - x_k \|.$$

Mais la suite x_k, $k \in \mathcal{J}$, a pour limite $\bar{x}$, aussi $\| x_j - x_k \| \leqslant r/(2K)$ pour tous les k et j suffisamment grands, de sorte que $f(x_j) \leqslant r/2$ pour j importants, ce qui contredit $f(x_j) \geqslant r$, $j \in \mathcal{J}$.

Ainsi, on a établi que $f(x_k)$ tend vers zéro. Soit $\bar{x}$ un point limite arbitraire, i.e. $x_k \to \bar{x}$, $k \in \mathcal{J}$, où $\mathcal{J}$ est une suite partielle d'indices. Etant donné la continuité de $f(x)$,

$$f(\bar{x}) = \lim_{k \in \mathcal{J}} f(x_k) = 0,$$

i.e. $\bar{x} \in \Omega$. Par ailleurs, $f_0(x_k) \leqslant f_0(x_*)$ et donc $f_0(\bar{x}) \leqslant f_0(x_*)$, d'où il découle de suite que $f_0(\bar{x}) = f_0(x_*)$ et que $\bar{x}$ est également une solution du problème (4.1). Le théorème est démontré.

2. Aspects numériques. L'algorithme de plan sécant exige qu'on résolve à chaque pas le problème de programmation linéaire : minimiser $f_0(x) = (c, x)$ dans les contraintes

$$(a_i, x) - b_i \leqslant 0, \quad i = -l, \ldots, k. \tag{4.6}$$

Ainsi, le problème augmente de volume d'un pas à l'autre. On utilise de même toujours plus de mémoires pour conserver les vecteurs a_k. Le problème (4.6) se résout plus simplement si l'on passe à son dual qui s'écrit en l'occurrence: maximiser $-\sum_{i=-l}^{k} u^i b_i$ dans les contraintes

$$\sum_{i=-l}^{k} u^i a_i + c = 0, \quad u^i \geqslant 0, \quad i = -l, \ldots, k.$$

En l'abordant par la méthode du simplexe la solution de ce problème sert de solution admissible pour le problème suivant. C'en est également une bonne approximation si bien que le nouveau problème est résolu au bout d'un nombre faible d'itérations.

On a à chaque pas à calculer a_{k+1} qui est un vecteur d'appui à $f(x)$ en x_k. Rappelons (voir ch. I) que si $f(x)$ est une fonction dérivable, alors $a_{k+1} = f'(x_k)$. Mais si $f(x) = \max_{1 \leqslant i \leqslant m} f_i(x)$, on prend comme a_{k+1} tout vecteur de la forme $\sum_{i \in \mathcal{J}(x_k)} \lambda_i f_i'(x_k)$, où $\lambda_i \geqslant 0$, $\sum_{i \in \mathcal{J}(x_k)} \lambda_i = 1$, $\mathcal{J}(x_k) = \{i : f_i(x_k) = f(x_k), 1 \leqslant i \leqslant m\}$; en particulier, on peut poser $a_{k+1} = f_i'(x_k)$, avec i n'importe quel indice de $\mathcal{J}(x_k)$.

Les règles que nous venons de citer découlent du § 2, ch. I.

3. Remarques terminales. Notre exposé de la méthode du plan sécant suit l'article de J. E. Kelley [81]. On connaît plusieurs variantes de cette technique (consulter p.ex. E. Lévitine et B. Poliak [22]). Il paraît que toutes ces mises en œuvre n'améliorent pas la propriété qui nous préoccupe, à savoir la vitesse de convergence, qui n'est pas exactement évaluée pour la méthode exposée mais qui à en juger d'après Kelley n'est même pas géométrique.

§ 5. Méthode de linéarisation

Nous allons considérer une méthode de résolution du problème général de programmation mathématique sans faire des hypothèses sur la convexité des fonctions. Son trait distinctif est de prendre en considération des contraintes non linéaires de type égalité, ce qui constitue une difficulté insurmontable pour la plupart des autres méthodes.

Soit à minimiser une fonction $f_0(x)$, $x \in E^n$, sous les contraintes

$$f_i(x) \leqslant 0, \quad i \in \mathcal{J}^-, \quad f_i(x) = 0, \quad i \in \mathcal{J}^0, \tag{5.1}$$

avec $\mathcal{J}^-$ et $\mathcal{J}^0$ deux ensembles finis d'indices. Supposons que toutes les fonctions $f_i(x)$ admettent des dérivées continues. (Les conditions

du problème seront précisées plus bas.) Par linéarisation de $f_i(x)$ en x remplaçons au point x_0 toutes les contraintes (5.1) et $f_0(x)$ par leurs homologues linéaires. On obtient un problème de programmation linéaire. Il serait naturel d'imiter la méthode de Newton pour des systèmes d'équations non linéaires et de prendre pour approximation suivante la solution du problème linéarisé. Cette voie telle quelle ne conduit malheureusement pas au but parce que le problème auxiliaire de la programmation linéaire ne possède en général pas de solution. On impose donc certaines contraintes à l'accroissement du vecteur x au point x_0 de façon que la solution du problème linéarisé en ce point ne s'en éloigne pas trop et reste dans un voisinage de x_0 où la linéarisation est valable. Nous le ferons ci-dessous par adjonction d'un terme quadratique à la fonction économique linéarisée.

Notons que toute égalité $f_i(x) = 0$ équivaut à deux inégalités

$$f_i(x) \leqslant 0, \quad -f_i(x) \leqslant 0.$$

On se bornera donc au cas avec des contraintes inégalité. Cet artifice allège au moins la motivation théorique de l'algorithme bien que le doublement du nombre d'inégalités risque d'alourdir le calcul. Plus loin nous justifierons théoriquement l'algorithme pour le problème de minimisation de $f_0(x)$ sous les contraintes

$$f_i(x) \leqslant 0, \ i \in \mathcal{J}. \tag{5.2}$$

L'algorithme modifié pour le problème général (5.1) sera étudié séparément.

Ainsi, nous étudierons, sans réduire la généralité, l'algorithme pour le problème (5.2). Il est clair qu'on peut toujours supposer l'existence d'une inégalité (5.2) triviale : $0 \leqslant 0$. On admettra donc qu'il y a, parmi $f_i(x)$, $i \in \mathcal{J}$, une fonction identiquement nulle : $f_i(x) \equiv 0$.

1. Hypothèses fondamentales. Posons

$$\begin{gathered} F(x) = \max_{i \in \mathcal{J}} f_i(x), \\ \mathcal{J}_\delta(x) = \{i \in \mathcal{J} : f_i(x) \geqslant F(x) - \delta\}, \quad \delta \geqslant 0, \end{gathered} \tag{5.3}$$

Par hypothèse faite plus haut, $F(x) \geqslant 0$ quel que soit x. Supposons qu'il existe des constantes $N > 0$, $\delta > 0$, telles que

a) l'ensemble

$$\Omega_N = \{x : f_0(x) + NF(x) \leqslant C_0\}, \ C_0 = f_0(x_0) + NF(x_0),$$

soit borné ;

b) les gradients des fonctions $f_i(x)$, $i \in \{0\} \cup \mathcal{J}$, vérifient dans Ω_N la condition de Lipschitz, i.e.

$$\| f_i'(x_1) - f_i'(x_2) \| \leqslant L \| x_1 - x_2 \|;$$

c) le problème de programmation quadratique

$$\min (f_0'(x), p) + \frac{1}{2} \| p \|^2, \tag{5.4}$$

$$(f_i'(x), p) + f_i(x) \leqslant 0, \quad i \in \mathcal{J}_\delta(x),$$

soit résoluble par rapport à $p \in E^n$ pour tout $x \in \Omega_N$ et il existe des multiplicateurs de Lagrange $u^i(x)$, $i \in \mathcal{J}_\delta(x)$, tels que $\sum_{i \in \mathcal{J}_\delta(x)} u^i(x) \leqslant N$. Ici et partout dans ce paragraphe $\| p \|$ désigne la norme euclidienne du vecteur p.

La solution du problème (5.4) sera désignée par $p(x)$ et les multiplicateurs de Lagrange par $u^i(x)$, $i \in \mathcal{J}_\delta(x)$.

2. Formulation de l'algorithme. Soit l'approximation initiale x_0 et ε choisi, $0 < \varepsilon < 1$. Supposons que le point x_k est déjà obtenu. La construction de l'approximation suivante se fait en deux étapes.

1) Résolvons le problème (5.4) pour $x = x_k$ et trouvons sa solution, à savoir le vecteur $p_k \equiv p(x_k)$.

2) Trouvons la première valeur de $i = 0, 1, \ldots$, satisfaisant à l'inégalité

$$f\left(x_k + \frac{1}{2^i} p_k\right) + NF\left(x_k + \frac{1}{2^i} p_k\right) \leqslant f(x_k) + NF(x_k) - \frac{1}{2^i} \varepsilon \| p_k \|^2.$$

Si cette inégalité a lieu la première fois pour $i = i_0$, on pose $\alpha_k = 2^{-i_0}$, $x_{k+1} = x_k + \alpha_k p_k$.

Ainsi, on a à chaque pas

$$f(x_{k+1}) + NF(x_{k+1}) \leqslant f(x_k) + NF(x_k) - \alpha_k \varepsilon \| p_k \|^2. \tag{5.5}$$

3. Convergence de l'algorithme. Montrons qu'à chaque itération le choix du pas α_k s'effectue en un nombre fini de bipartitions de l'unité et justifions la convergence de l'algorithme.

Les résultats du ch. I, § 3 entraînent que $p(x)$ est solution du problème (5.4) si et seulement si il existe des $u^i(x) \geqslant 0$, $i \in \mathcal{J}_\delta(x)$, tels que

$$f_0'(x) + p(x) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i'(x) = 0, \tag{5.6}$$

$$u^i(x)((f_i'(x), p(x)) + f_i(x)) = 0, \quad i \in \mathcal{J}_\delta(x).$$

Aussi

$$(f_0'(x), p(x)) = -\sum_{i \in \mathcal{J}_\delta(x)} u^i(x)(f_i'(x), p(x)) - \| p(x) \|^2 =$$

$$= \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i(x) - \| p(x) \|^2. \tag{5.7}$$

LEMME 5.1. *Pour que le point x vérifie les inégalités* (5.2) *et les conditions nécessaires de minimum de $f_0(x)$ avec les contraintes* (5.2), *il faut et il suffit qu'on ait l'égalité $p(x) = 0$.*

DÉMONSTRATION. Supposons que x vérifie (5.2) et qu'on a en ce point les conditions nécessaires de minimum de $f_0(x)$. Il existe alors des nombres $u^i \geqslant 0$, $i \in \mathcal{J}$, tels que

$$f_0'(x) = \sum_{i \in \mathcal{J}} u^i f_i'(x) = 0, \quad u^i f_i(x) = 0, \quad i \in \mathcal{J}. \tag{5.8}$$

Si x satisfait à (5.2), alors $F(x) = 0$ et $\mathcal{J}_0(x)$ se confond donc avec l'ensemble des i pour lesquels $f_i(x) = 0$. En vertu de la seconde relation (5.8) on a de plus $u^i = 0$ si $f_i(x) < 0$, i.e. si $i \notin \mathcal{J}_0(x)$. Vu que $\mathcal{J}_\delta(x) \supseteq \mathcal{J}_0(x)$, (5.8) se récrit donc:

$$f_0'(x) + \sum_{i \in \mathcal{J}_\delta(x)} u^i f_i'(x) = 0, \quad u^i f_i(x) = 0, \quad i \in \mathcal{J}_\delta(x).$$

Mais la comparaison de (5.6) et des dernières relations montre que le vecteur $p = 0$ est solution du problème (5.4) car toutes les contraintes (5.4) sont vérifiées pour cette valeur de p (du moment que (5.2) le sont), et la justesse des relations (5.6) pour $p = 0$ constitue une condition nécessaire et suffisante pour que le vecteur $p = 0$ soit solution de (5.4).

Soit maintenant $p(x) = 0$, autrement dit, les contraintes du problème (5.4) sont satisfaites pour $p = 0$, i.e. $f_i(x) \leqslant 0$, $i \in \mathcal{J}_\delta(x)$. Comme on a, pour $i \notin \mathcal{J}_\delta(x)$,

$$f_i(x) \leqslant F(x) - \delta \leqslant f_j(x) \leqslant 0,$$

avec $j \in J_\delta(x)$, le point x vérifie toutes les contraintes (5.2). De plus, quand $p = 0$, les relations (5.6) deviennent (5.8) si l'on pose $u^i = 0$, $i \notin \mathcal{J}_\delta(x)$. Ainsi, les conditions nécessaires de minimum de $f_0(x)$ sous les contraintes (5.2) sont également remplies, ce qui achève la démonstration.

Evaluons la variation que présenteront toutes les fonctions du problème lorsqu'on se déplacera du point x_k dans la direction p_k.

Pour $i \in \mathcal{J}_\delta(x_k)$, on a, à l'aide de la formule de Taylor,

$$f_i(x_k + \alpha p_k) = f_i(x_k) + \alpha(p_k, f_i'(x_k)) + \alpha(p_k, f_i'(\theta_i) - f_i'(x_k)),$$

où $\theta_i = x_k + \alpha \xi_i p_k$, $0 \leqslant \xi_i \leqslant 1$. p_k étant solution de (5.4) quand $x = x_k$, on a

$$\begin{aligned} f_i(x_k + \alpha p_k) &\leqslant f_i(x_k) - \alpha f_i(x_k) + \alpha^2 \| p_k \|^2 L \leqslant \\ &\leqslant (1 - \alpha) f_i(x_k) + \alpha^2 \| p_k \|^2 L, \end{aligned} \tag{5.9}$$

formule déduite en se rappelant que les gradients de $f_i(x)$ vérifient la condition de Lipschitz.

Pour $i \notin \mathcal{J}_\delta(x_k)$, on a

$$f_i(x_k+\alpha p_k)=f_i(x_k)+\alpha(p_k, f_i'(\theta_i))\leqslant F(x_k)-\delta+\alpha K\|p_k\|, \quad (5.10)$$

avec K bornant $\|f_i'(x)\|$ dans Ω_N.

Du moment que

$$(1-\alpha)F(x_k)\geqslant F(x_k)-\delta+\alpha K\|p_k\|$$

pour α tels que $\alpha\leqslant 1$,

$$0\leqslant\alpha\leqslant\frac{\delta}{F(x_k)+K\|p_k\|}, \quad (5.11)$$

(5.9) et (5.10) entraînent que, quel que soit i, on a l'inégalité

$$f_i(x_k+\alpha p_k)\leqslant(1-\alpha)F(x_k)+\alpha^2 L\|p_k\|^2 \quad (5.12)$$

à condition que α vérifie (5.11).

Par analogie avec les estimations précédentes

$$f_0(x_k+\alpha p_k)=f_0(x_k)+\alpha(p_k, f_0'(x_k))+\alpha(p_k, f_0'(\theta_0)-f_0(x_k)),$$
$$\theta_0=x_k+\alpha\xi_0 p_k, \quad 0\leqslant\xi_0\leqslant 1.$$

Utilisons (5.7) et la condition de Lipschitz pour les gradients, il vient

$$f_0(x_k+\alpha p_k)\leqslant f_0(x_k)+\alpha\Big(\sum_{i\in\mathcal{J}_\delta(x_k)}u^i(x_k)f_i(x_k)\Big)-\alpha\|p_k\|^2+\alpha^2 L\|p_k\|^2.$$

Cette relation et (5.12) entraînent

$$f_0(x_k+\alpha p_k)+NF(x_k+\alpha p_k)\leqslant f_0(x_k)+NF(x_k)+$$
$$+\alpha\Big(\sum_{i\in\mathcal{J}_\delta(x_k)}u^i(x_k)f_i(x_k)-NF(x_k)\Big)-\alpha\|p_k\|^2+\alpha^2(N+1)L\|p_k\|^2. \quad (5.13)$$

Rappelons que $u^i(x_k)\geqslant 0$, $F(x_k)\geqslant 0$ et

$$\sum_{i\in\mathcal{J}_\delta(x_k)}u^i(x_k)\leqslant N.$$

Donc

$$\sum_{i\in\mathcal{J}_\delta(x_k)}u^i(x_k)f_i(x_k)-NF(x_k)\leqslant 0.$$

(5.13) se récrit alors:

$$f_0(x_k+\alpha p_k)+NF(x_k+\alpha p_k)\leqslant$$
$$\leqslant f_0(x_k)+NF(x_k)-\alpha\|p_k\|^2(1-\alpha(N+1)L)$$

ou (si

$$0\leqslant\alpha\leqslant\frac{1-\varepsilon}{(N+1)L}\Big) \quad (5.14)$$

$$f_0(x_k+\alpha p_k)+NF(x_k+\alpha p_k)\leqslant f_0(x_k)+NF(x_k)-\alpha\varepsilon\|p_k\|^2. \quad (5.15)$$

Si

$$0 \leqslant \alpha \leqslant \bar{\alpha}_k,$$

$$\bar{\alpha}_k = \min\left(1, \frac{\delta}{F(x_k) + K \| p_k \|}, \frac{1-\varepsilon}{(N+1)L}\right),$$

on a donc l'inégalité (5.15).

Cela signifie que (5.5) a lieu après un nombre fini d'essais $\alpha = 2^{-i}$, $i = 0, 1, \ldots$, et qu'on a

$$\alpha_k > \frac{1}{2}\bar{\alpha}_k. \tag{5.16}$$

Démontrons un théorème sur la convergence du processus.

THEOREME 5.1. *Sous les hypothèses du* n° 1, *le processus jouit des propriétés suivantes*:

a) $F(x_k) \to 0$ *pour* $k \to \infty$;

b) *tout point limite* x_* *de la suite* x_k, $k = 0, 1, \ldots$, *vérifie les inégalités* (5.2) *et les conditions nécessaires de minimum de* $f_0(x)$ *sous les contraintes* (5.2).

REMARQUE. Dire que $F(x_k)$ tend vers zéro c'est dire que la suite x_k satisfait toujours plus exactement aux contraintes (5.2).

DEMONSTRATION. La fonction $f_0(x) + NF(x)$ décroissant à chaque pas en vertu de (5.15), tous les points de x_k appartiennent au domaine Ω_N. Comme ce dernier est compact, $f_0(x) + NF(x)$ est bornée sur cet ensemble parce qu'elle est continue. D'où

$$\alpha_k \| p_k \|^2 \to 0 \tag{5.17}$$

pour $k \to \infty$, car dans le cas contraire $f_0(x) + NF(x)$ décroît indéfiniment suivant la suite x_k.

Démontrons que $p_k \to 0$. En effet, s'il n'en était pas ainsi, (5.17) entraînerait $\alpha_k \to 0$ suivant une suite partielle d'indices k. Il résulte alors de (5.16) et de l'expression de $\bar{\alpha}_k$ que pour k grands on a

$$\alpha_k \geqslant \frac{1}{2}\bar{\alpha}_k = \frac{1}{2}\frac{\delta}{F(x_k) + K \| p_k \|}.$$

Le second membre de la dernière inégalité doit donc tendre vers zéro. Comme $F(x)$ est une fonction continue sur l'ensemble compact Ω_N elle est majorée et l'expression $\frac{\delta}{F(x_k) + K \| p_k \|}$ ne tend vers zéro que si $\| p_k \| \to +\infty$. Or, on obtient à partir de (5.6)

$$\| p(x_k) \| = \left\| f_0'(x_k) + \sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k) f_i'(x_k) \right\| \leqslant K(N+1).$$

Ainsi, l'hypothèse de $p_k \not\to 0$ nous a conduit à une contradiction.

On a par définition de p_k

$$(f_i'(x_k), p_k) + f_i(x_k)) \leqslant 0, \quad i \in \mathcal{J}_\delta(x_k).$$

Donc

$$f_i(x_k) \leqslant -(f_i'(x_k), p_k) \leqslant K \| p_k \|, \quad i \in \mathcal{J}_\delta(x_k).$$

Mais $f_j(x_k) \leqslant f_i(x_k)$, $j \notin \mathcal{J}_\delta(x_k)$, $i \in \mathcal{J}_\delta(x_k)$. D'où

$$F(x_k) = \max_{i \in \mathcal{J}} f_i(x_k) \leqslant K \| p_k \|.$$

Ainsi, $F(x_k) \to 0$ pour $k \to \infty$ car $F(x_k) \geqslant 0$. Posons ensuite $u^i(x) = 0$, $i \notin \mathcal{J}_\delta(x)$. (5.6) se récrit alors suivant la suite x_k:

$$\begin{gathered} f_0'(x_k) + p_k + \sum_{i \in \mathcal{J}} u^i(x_k) f_i'(x_k) = 0, \\ u^i(x_k)((f_i'(x_k), p_k) + f_i(x_k)) = 0, \ i \in \mathcal{J}. \end{gathered} \tag{5.18}$$

Soit x_* un point limite de la suite x_k. Du moment que $x_k \in \Omega_N$ et Ω_N est compact, il existe toujours de tels points. On peut estimer, sans restreindre la généralité, que $x_k \to x_*$. Etant donné que $u^i(x) \geqslant 0$, $i \in \mathcal{J}$, et que leur somme est bornée, on admet de plus que $u^i(x_k) \to u^i$ quand $k \to \infty$.

Passons à la limite dans (5.18), il vient

$$f_0'(x_*) + \sum_{i \in \mathcal{J}} u^i f_i'(x_*) = 0, \ u^i f_i(x_*) = 0, \ i \in \mathcal{J}.$$

De plus, $u^i \geqslant 0$ car $u^i(x_k) \geqslant 0$, et le point x_* vérifie toutes les contraintes (5.2) parce que $f_i(x_k) \leqslant F(x_k)$ et $F(x_k) \to 0$, d'où, par passage à la limite, $f_i(x_*) \leqslant 0$. Cela démontre les conditions nécessaires de minimum en x_* et, par là même, le théorème 5.1.

Conséquence. *Si les conditions nécessaires de minimum ne sont remplies qu'au point qui réalise le minimum de la fonction la suite générée par l'algorithme converge dans les contraintes* (5.2) *vers le point de minimum de* $f_0(x)$.

En effet, seul un point de minimum peut être alors, selon le théorème 5.1, l'unique point limite de la suite x_k.

4. Aspects numériques. C'est la résolution de (5.4), problème de programmation quadratique, qui constitue l'opération principale exigeant de nombreux calculs à chaque pas. En optant pour telle méthode de résolution on se rappellera qu'il s'agit du problème auxiliaire qui doit être résolu en un nombre fini de pas. Etant donné qu'on ne connaît en général pas d'avance la constante N, il y a intérêt, en abordant (5.4), à obtenir les multiplicateurs de Lagrange $u^i(x)$ correspondants. A la lumière de ces circonstances on fera

bon de passer au dual de (5.4) et de se servir de la méthode du gradient conjugué exposée au n° 5, § 1 du présent chapitre.

Construisons ce problème dual. Conformément au ch. I, § 3, sa fonction économique s'écrit

$$\varphi(u) = \min_p \left[(f_0'(x), p) + \frac{1}{2} \| p \|^2 + \sum_{i \in \mathcal{J}_\delta(x)} u^i ((f_i'(x), p) + f_i(x)) \right]. \tag{5.19}$$

Egalons à zéro les dérivées par rapport à p du second membre, on constate que le minimum est réalisé pour

$$p = -f_0'(x) - \sum_{i \in \mathcal{J}_\delta(x)} u^i f_i'(x). \tag{5.20}$$

Ainsi, le point p est défini de façon unique par le vecteur u de composantes u^i, $i \in \mathcal{J}_\delta(x)$.

Portons (5.20) dans le second membre de (5.19), il vient

$$\varphi(u) = -\frac{1}{2} \left\| f_0'(x) + \sum_{i \in \mathcal{J}_\delta(x)} u^i f_i'(x) \right\|^2 + \sum_{i \in \mathcal{J}_\delta(x)} u^i f_i(x). \tag{5.21}$$

La fonction économique du problème dual est donc calculée, et le problème même consiste à maximiser $\varphi(u)$ sous les contraintes $u^i \geqslant 0$, $i \in \mathcal{J}_\delta(x)$.

Nous avons donc abouti à un problème de maximisation d'une forme quadratique avec des contraintes simples, problème qui se prête avec succès à la méthode du gradient conjugué (n° 5, § 1, ch. III).

La résolution nous donne les multiplicateurs de Lagrange $u^i(x)$, solution du dual, et, selon le § 3, ch. I, par substitution de $u^i(x)$ dans (5.20) on obtient le vecteur $p(x)$, solution du primal.

Une autre question à éclaircir est le choix des constantes N et δ. La quantité N est en général inconnue. En la prenant trop grande on risque, en vertu de la formule (5.14), de fractionner considérablement le pas. Il y a donc intérêt à l'évaluer au cours de l'algorithme. Si on a par exemple à un pas

$$N \leqslant \sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k),$$

on change N :

$$N = 2 \sum_{i \in \mathcal{J}_\delta(x_k)} u^i(x_k). \tag{5.22}$$

L'expérience atteste le bien-fondé de cette correction. Il est clair théoriquement que si x_k est suffisamment proche d'un point limite, $u^i(x_k)$ sont voisins, dans le cas régulier, des multiplicateurs de

Lagrange en x_*, solution du problème, et avec la formule (5.22) on va donc réussir. Plus loin nous parlerons en détail du comportement des multiplicateurs $u^i(x_k)$.

En ce qui concerne δ, il faut le réduire si le problème auxiliaire (5.4) se trouve impossible à un pas.

Donnons maintenant les conditions garanties d'existence de N et δ. En fait, ces constantes existent dans une classe de problèmes infiniment plus vaste.

THÉORÈME 5.2. *Si toutes les fonctions $f_0(x)$, $f_i(x)$, $i \in \mathcal{J}$, sont convexes, il existe un point $\bar{x}$ tel que*

$$f_i(\bar{x}) < 0, \quad i \in \mathcal{J},$$

$f_0(x)$ tendant vers $+\infty$ pour $x \to +\infty$ et le point x_0 vérifiant les contraintes (5.2), *quel que soit $\delta > 0$, les multiplicateurs $u^i(x)$, $i \in \mathcal{J}_\delta(x)$, sont alors majorés sur l'ensemble Ω_N pour N suffisamment grands et Ω_N est compact.*

DÉMONSTRATION. Rappelons que

$$\Omega_N = \{x\colon f_0(x) + NF(x) \leqslant f_0(x_0) + NF(x_0)\}.$$

Cette égalité et la continuité de $f_0(x)$ et de $F(x)$ entraînent que Ω_N est fermé. D'autre part Ω_N est borné parce que $f_0(x) \to +\infty$ pour $x \to +\infty$ par hypothèse et donc

$$f_0(x) + NF(x) > f_0(x_0) + NF(x_0)$$

avec tous les x suffisamment importants en norme. Poursuivons. Comme x_0 vérifie (5.2), $F(x_0) = 0$. Aussi $\Omega_N \subset \Omega_0$ quels que soient N. Il résulte en effet de $x \in \Omega_N$

$$f_0(x) \leqslant f_0(x) + NF(x) \leqslant f_0(x_0) + NF(x_0) = f_0(x_0),$$

i.e. $f_0(x) \leqslant f_0(x_0)$. La compacité de Ω_0 est également évidente en vertu de l'hypothèse du théorème.

Puisque toutes les $f_i(x)$ sont convexes, on a

$$f_i(x) + (f_i'(x), \bar{x} - x) \leqslant f_i(\bar{x}) < 0. \tag{5.23}$$

Le système de contraintes du problème (5.4) est donc compatible pour tout $\delta > 0$ du moment qu'il est vérifié par le vecteur $\bar{p} = \bar{x} - x$.

Soient $u^i(x)$ les multiplicateurs de Lagrange du problème (5.4). Selon le théorème du Kuhn et Tucker,

$$\frac{1}{2}\|p(x)\|^2 + (f_0'(x), p(x)) \leqslant$$

$$\leqslant \frac{1}{2}\|p\|^2 + (f_0'(x), p) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x)((f_i'(x), p) + f_i(x))$$

quel que soit p. En particulier, lorsque $p = \bar{p} \equiv \bar{x} - x$, on a en vertu de (5.23)

$$\frac{1}{2}\|p(x)\|^2 + (f_0'(x), p(x)) \leqslant$$

$$\leqslant \frac{1}{2}\|\bar{p}\|^2 + (f_0'(x), \bar{p}) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x)((f_i'(x), \bar{p}) + f_i(x)) \leqslant$$

$$\leqslant \frac{1}{2}\|\bar{p}\|^2 + (f_0'(x), \bar{p}) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i(\bar{x}) \leqslant$$

$$\leqslant \frac{1}{2}\|\bar{p}\|^2 + (f_0'(x), \bar{p}) + u^i(x) f_i(\bar{x}).$$

D'où ($f_i(\bar{x}) < 0$!),

$$u^i(x) \leqslant \frac{\left[\frac{1}{2}\|p(x)\|^2 + (f_0'(x), p(x))\right] - \left[\frac{1}{2}\|\bar{p}\|^2 + (f_0'(x), \bar{p})\right]}{f_i(\bar{x})}, \quad (5.24)$$

le numérateur contenant une quantité non positive parce que $p(x)$ est solution de (5.4) et $\bar{p}$ satisfaisant aux contraintes (5.4).

Montrons la borne dans Ω_0 de la fraction ci-dessus. En effet, les fonctions $f_i(x)$ admettant des dérivées continues, la quantité

$$\frac{1}{2}\|\bar{p}\|^2 + (f_0'(x), \bar{p}) = \frac{1}{2}\|\bar{x} - x\|^2 + (f_0'(x), \bar{x} - x)$$

est bornée dans le compact Ω_0. Aussi la quantité plus petite

$$\frac{1}{2}\|p(x)\|^2 + (f_0'(x), p(x))$$

y est majorée. Quant à sa minoration, on a

$$\frac{1}{2}\|p(x)\|^2 + (f_0'(x), p(x)) \geqslant \frac{1}{2}\|p(x)\|^2 - \|f_0'(x)\|\,\|p(x)\| \geqslant$$

$$\geqslant -\frac{1}{2}\|f_0'(x)\|^2,$$

i.e. pour $x \in \Omega_0$ la quantité étudiée est également minorée.

Ainsi, nous avons montré que les seconds membres de (5.24) sont majorés dans Ω_0, i.e. $u^i(x) \leqslant M$, $x \in \Omega_0$, il en découle de suite l'affirmation du théorème.

Si le problème primal est donc un problème de programmation convexe, l'algorithme travaille avec tout $\delta > 0$ à condition que le domaine admissible contienne le point intérieur.

5. Certaines généralisations. Nous avons déjà dit au début de ce paragraphe qu'en présence de contraintes de type égalité, i.e.

quand les contraintes sont de la forme (5.1), le problème se ramène à (5.2) en remplaçant chaque égalité par deux inégalités.

L'algorithme s'applique donc au problème général (5.1). Mais on n'oubliera pas que si pour un x

$$f_i(x) \geqslant F(x) - \delta \text{ et } -f_i(x) \geqslant F(x) - \delta,$$

avec $i \in \mathcal{J}_0$, le système (5.4) comprend alors deux inégalités

$$(f_i'(x), p) + f_i(x) \leqslant 0, \quad -(f_i'(x), p) - f_i(x) \leqslant 0, \tag{5.25}$$

qui équivalent à l'égalité

$$(f_i'(x), p) + f_i(x) = 0. \tag{5.26}$$

Il y a donc intérêt à en tenir compte en résolvant le problème auxiliaire et à substituer dans (5.4) une égalité (5.26) à deux inégalités de la forme (5.25). En passant au problème dual, cela détermine l'arbitraire quant au signe du multiplicateur u^i correspondant, ce qui n'exclut cependant pas l'algorithme de gradient conjugué (n° 5, § 1, ch. III).

Admettons que le problème primal est soumis, en plus de (5.2), à une contrainte définie par la condition que le point x appartient à un ensemble X de structure simple. Il serait bon dans ce cas que les approximations obtenues se trouvent dans X. Indiquons la modification correspondante de l'algorithme. Toujours sans restreindre la généralité, bornons-nous au cas de contraintes inégalité.

Ainsi, on demande de minimiser $f_0(x)$, $x \in E^n$, dans les contraintes

$$f_i(x) \leqslant 0, \quad i \in \mathcal{J}, \ x \in X, \tag{5.27}$$

où $\mathcal{J}$ est un ensemble fini d'indices et X un ensemble fermé convexe. On suppose l'existence d'un indice i tel que $f_i(x) \equiv 0$.

Admettons qu'il existe des constantes $N > 0$ et $\delta > 0$ vérifiant les conditions suivantes :

a) l'ensemble

$$\Omega_N = \{x\colon f_0(x) + NF(x) \leqslant C_0, \ x \in X\},$$
$$C_0 = f_0(x_0) + NF(x_0),$$

est borné et l'approximation initiale x_0 appartient à X ;

b) les gradients des fonctions $f_i(x)$, $i \in \{0\} \cup \mathcal{J}$, vérifient dans Ω_N la condition de Lipschitz, i.e.

$$\| f_i'(x_1) - f_i'(x_2) \| \leqslant L \| x_1 - x_2 \| ;$$

c) le problème

$$\min (f_0'(x), p) + \frac{1}{2} \| p \|^2, \tag{5.28}$$

$$(f_i'(x), p) + f_i(x) \leqslant 0, \ i \in \mathcal{J}_\delta(x), \ x + p \in X,$$

est résoluble par rapport à p pour n'importe quel $x \in \Omega_N$, et il existe des multiplicateurs de Lagrange $u^i(x)$, $i \in \mathcal{J}_\delta(x)$, tels que

$$\sum_{i \in \mathcal{J}_\delta(x)} u^i(x) \leqslant N.$$

Remarque. Rappelons que les multiplicateurs de Lagrange pour le problème (5.28) sont des nombres non négatifs satisfaisant aux conditions

$$(f_0'(x), p(x)) + (p(x), p(x)) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) [(f_i'(x), p(x)) + f_i(x)] \leqslant$$
$$\leqslant (f_0'(x), p) + (p(x), p) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) [(f_i'(x), p) + f_i(x)] \quad (5.29)$$

pour tous les p tels que

$$x + p \in X. \quad (5.30)$$

De plus,

$$u^i(x) [(f_i'(x), p(x)) + f_i(x)] = 0, \ i \in \mathcal{J}_\delta(x). \quad (5.31)$$

Ainsi, la condition c) suppose non seulement que le problème auxiliaire (5.28) possède une solution, mais aussi que les conditions nécessaires et suffisantes du théorème de Kuhn et Tucker sont réalisées au point de minimum $p = p(x)$.

L'algorithme de résolution de (5.27) est maintenant construit comme indiqué au n° 2 de ce paragraphe, à cette différence près qu'on prend pour p_k le vecteur $p(x_k)$, solution du nouveau problème auxiliaire (5.28).

Montrons la convergence de l'algorithme, i.e. la validité du théorème 5.1, et l'appartenance de x_k à X quels que soient k. La dernière affirmation implique en particulier la même propriété de tout point limite de la suite x_k. Vu que la démonstration de la convergence ne diffère de celle du théorème 5.1 que par certains détails il suffit d'en citer les éléments caractéristiques.

Primo, comme $x_k + p_k \in X$, avec X convexe, on a $x_k + \alpha p_k \in X$ pour tout α compris entre 0 et 1. Si $x_k \in X$, on a donc $x_{k+1} \in X$. Or, x_0 appartenant à X par hypothèse, il en est de même de toute la suite $\{x_k\}_{k=0}^{\infty}$.

Secundo, on obtient de (5.29)-(5.31) pour $p = 0$

$$(f_0'(x), p(x)) + \| p(x) \|^2 \leqslant \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i(x),$$

i.e.

$$(f_0'(x), p(x)) \leqslant \sum_{i \in \mathcal{J}_\delta(x)} u^i(x) f_i(x) - \| p(x) \|^2. \quad (5.32)$$

Cette inégalité remplace (5.7) utilisée dans l'évaluation (5.13), les autres calculs effectués en vue d'obtenir des estimations restant sans modification.

Enfin, si $p(x_*) = 0$ en x_*, il découle de (5.29)-(5.31) les conditions suivantes:

$$\begin{gathered}(f_0'(x_*), p) + \sum_{i \in \mathcal{J}_\delta(x)} u^i(x_*)(f_i'(x_*), p) \geqslant 0, \\ x_* + p \in X, \quad u^i(x_*) f_i(x_*) = 0, \quad i \in \mathcal{J}_\delta(x_*).\end{gathered} \tag{5.33}$$

De plus (5.28) entraîne alors

$$f_i(x_*) \leqslant 0, \quad i \in \mathcal{J}_\delta(x_*), \ x_* \in X,$$

et il est évident que

$$f_i(x_*) < 0, \quad i \notin \mathcal{J}_\delta(x_*).$$

Ainsi, le point x_* vérifie toutes les contraintes (5.27) et les conditions (5.33) montrent qu'on a en ce point les conditions nécessaires d'extrémum.

Nous avons montré une fois de plus que, si $p(x_*) = 0$, le point x_* remplit les conditions nécessaires d'extrémum. La réciproque se démontre sans peine de sorte que $p(x) = 0$ est une condition nécessaire et suffisante pour que le point x soit soupçonné de réaliser l'extrémum de la fonction.

Comme pour le théorème 5.1, on démontre par passage des relations (5.29)-(5.31) remplies aux points x_k aux relations (5.33) qui le sont au point limite, que tout point limite x_* de la suite x_k, $k = 0, 1, \ldots$, vérifie les conditions nécessaires d'extrémum.

6. Problème de programmation linéaire. Supposons maintenant toutes les fonctions $f_0(x)$, $f_i(x)$, $i \in \mathcal{J}$, de (5.2) linéaires, si bien qu'on est conduit au problème de programmation linéaire. Quoique l'algorithme exposé plus haut soit le plus intéressant pour le cas non linéaire, son application au cas linéaire a un sens. En particulier, si l'ensemble d'indices $\mathcal{J}$ se compose de nombreux éléments, on a un problème de programmation linéaire avec de nombreuses contraintes. D'autre part, pour δ petit, le problème auxiliaire (5.4) ne présente que peu de contraintes si bien que le problème général se ramène à la résolution d'une suite de problèmes plus simples. A la différence de la méthode du simplexe, la technique proposée ne conduit pas à l'accumulation de l'erreur de calcul parce qu'elle ne transforme pas à chaque pas la matrice initiale des contraintes.

S'agissant du problème de programmation linéaire, les conditions fondamentales a), c) (la condition b) est vérifiée automatiquement) imposées à la convergence s'avèrent trop sévères. Nous ne nous arrêterons pas sur les conditions de convergence pour le problème de programmation linéaire vu que notre but essentiel est de construire l'algorithme pour le cas non linéaire. On montrera plus loin qu'au moins dans a), c) pour le problème de programmation

linéaire il y a convergence au bout d'un nombre fini de pas. Ce fait caractérisera dans une certaine mesure la vitesse de convergence de l'algorithme.

THÉORÈME 5.3. *Soient, sous les hypothèses* a), c) *du* n° 1, $f_0(x)$, $f_i(x)$ *les fonctions définissant le problème* (5.2) *qui sont toutes de la forme*

$$f_i(x) = (a_i, x) - b_i.$$

L'algorithme du n° 2 *converge alors après un nombre fini de pas.*

DÉMONSTRATION. Notons pour commencer que dans notre cas le pas α_k vaut 1 pour des k suffisamment grands. En effet, toutes les $f_i(x)$ étant linéaires, la constante de Lipschitz L est nulle. Il résulte donc de la formule pour $\bar{\alpha}_k$ (n° 3)

$$\bar{\alpha}_k = \min\left(1, \frac{\delta}{F(x_k)+K\|p_k\|}, \frac{1-\varepsilon}{(N+1)L}\right) = \\ = \min\left(1, \frac{\delta}{F(x_k)+K\|p_k\|}\right). \tag{5.34}$$

Or, on a démontré plus haut que $F(x) \to 0$, $\|p_k\| \to 0$. On a donc, pour des k suffisamment grands, $\frac{\delta}{F(x_k)+K\|p_k\|} \geqslant 1$, et $\bar{\alpha}_k = 1$. Mais $\bar{\alpha}_k$ est construit de façon que l'inégalité (5.15) ait lieu quand $\alpha_k = \bar{\alpha}_k$. Comme le choix de α_k à chaque itération débute par la bipartition de $\alpha = 1$, il en résulte que l'inégalité (5.5) déterminant ce choix est vérifiée immédiatement sans qu'on ait besoin de fractionnements supplémentaires, et le pas α_k vaut 1.

Soit maintenant x_* un point limite de la suite x_k générée par l'algorithme. On sait qu'il est solution du problème (5.2) vu qu'il vérifie toutes les contraintes de celui-ci et qu'on a en ce point, selon le théorème 5.1, les conditions nécessaires de minimum qui s'avèrent également suffisantes dans le cas considéré du problème de programmation linéaire.

Posons

$$\mathcal{J}_0(x_*) = \{i \in \mathcal{J} : f_i(x_*) = 0\}. \tag{5.35}$$

Alors $f_i(x_*) < 0$ pour $i \notin \mathcal{J}_0(x_*)$, de sorte que

$$\varepsilon_0 = \max_{i \notin \mathcal{J}_0(x_*)} f_i(x_*) < 0. \tag{5.36}$$

Afin de simplifier les notations, nous supposerons, sans restreindre la généralité, que toute la suite x_k converge vers x_*.

Considérons le problème auxiliaire (5.4) aux points de la suite x_k :

$$\begin{gathered} \min\,(f_0'(x_k), p) + \frac{1}{2}\|p\|^2, \\ (f_i'(x_k), p) + f_i(x_k) \leqslant 0, \quad i \in \mathcal{J}_0(x); \end{gathered} \tag{5.37}$$

$p_k = p(x_k)$ en est solution. Désignons par u_k^i, $k \in \mathcal{J}_\delta(x_k)$, les multiplicateurs de Lagrange correspondants, de sorte que

$$u_k^i\,[(f_i'(x_k), p_k) + f_i(x_k)] = 0. \tag{5.38}$$

Montrons l'inclusion $\mathcal{J}_0(x_*) \subset \mathcal{J}_\delta(x_k)$ pour tous les k suffisamment grands. En effet, si $i \notin \mathcal{J}_\delta(x_k)$, alors

$$f_i(x_k) < F(x_k) - \delta,$$

et par passage à la limite par rapport à k on obtient, compte tenu de la tendance de $F(x_k)$ vers zéro, que $f_i(x_*) \leqslant -\delta$, ce qui contredirait $i \in \mathcal{J}_0(x_*)$.

Notons

$$\tilde{\mathcal{J}}(x_k) = \{i \in \mathcal{J}_\delta(x_k) : u_k^i > 0\}.$$

Nous disons que pour k importants

$$\tilde{\mathcal{J}}(x_k) \subset \mathcal{J}_0(x_*). \tag{5.39}$$

En effet, si $i \notin \mathcal{J}_0(x_*)$ alors $f_i(x_*) \leqslant \varepsilon_0$. Comme $p_k \to 0$, $f_i'(x_k)$ sont bornées et $x_k \to x_*$, on a pour k grands

$$|(f_i'(x_k), p_k)| \leqslant -\frac{\varepsilon_0}{4}, \quad f_i(x_k) \leqslant \frac{\varepsilon_0}{2},$$

et donc

$$(f_i'(x_k), p_k) + f_i(x_k) \leqslant \frac{\varepsilon_0}{4} < 0.$$

Donc, si $u_k^i > 0$, on a

$$u_k^i\,[(f_i'(x_k), p_k) + f_i(x_k)] < 0,$$

ce qui contredit (5.38).

Remarque. Les développements ci-dessus n'ont pas utilisé la linéarité de $f_i(x)$ si bien que les affirmations sur les inclusions $\mathcal{J}_0(x_*) \subset \mathcal{J}_\delta(x_k)$ et $\tilde{\mathcal{J}}(x_k) \subset \mathcal{J}_0(x_*)$ sont vraies dans le cas général d'un problème non linéaire et nous nous en servirons dans la suite.

Le dual du problème auxiliaire (5.37) consiste en la maximisation de la fonction (5.21) sous les contraintes $u^i \geqslant 0$, $i \in \mathcal{J}_\delta(x_k)$ (voir n° 4 du présent paragraphe). Ceci étant, les multiplicateurs de Lagrange u_k^i sont solution du problème dual et il y a égalité des valeurs optimales du primal et du dual, i.e.

$$(f_0'(x_k), p_k) + \frac{1}{2}\|p_k\|^2 =$$

$$= -\frac{1}{2}\|f_0'(x_k) + \sum_{i \in \mathcal{J}_\delta(x_k)} u_k^i f_i'(x_k)\|^2 + \sum_{i \in \mathcal{J}_\delta(x_k)} u_k^i f_i(x_k).$$

Comme $p_k \to 0$, le premier membre de la dernière relation tend vers zéro et donc

$$-\frac{1}{2}\left\| f_0'(x_k) + \sum_{i \in \mathcal{J}_\delta(x_k)} u_k^i f_i'(x_k) \right\|^2 + \sum_{i \in \mathcal{J}_\delta(x_k)} u_k^i f_i(x_k) \to 0. \quad (5.40)$$

Notons que u_k^i n'est > 0 que si $i \in \tilde{\mathcal{J}}(x_k)$. De plus

$$f_i(x) = (a_i, x) - b_i, \ i \in \{0\} \cup \mathcal{J},$$

de sorte que $f_i'(x) = a_i$ et est indépendant de x. Aussi (5.40) se récrit

$$-\frac{1}{2}\left\| a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_k^i a_i \right\|^2 \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_k^i f_i(x_k) \to 0.$$

Mais nous savons que $\tilde{\mathcal{J}}(x_k) \subset \mathcal{J}_0(x_*)$, c'est pourquoi $f_i(x_k) \to \to f_i(x_*) = 0$, car $f_i(x_*) = 0$ par définition pour $i \in \mathcal{J}_0(x_*)$. Donc

$$-\frac{1}{2}\left\| a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_k^i a_i \right\|^2 \to 0.$$

Mais

$$-\frac{1}{2}\left\| a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_k^i a_i \right\|^2 \leqslant \max_{u^i \geqslant 0,\ i \in \tilde{\mathcal{J}}(x_k)} -\frac{1}{2}\left\| a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u^i a_i \right\|^2 \leqslant 0. \quad (5.41)$$

Posons

$$\omega(\mathcal{Y}) = \max_{u^i \geqslant 0,\ i \in \mathcal{Y}} -\left\| a_0 + \sum_{i \in \mathcal{Y}} u^i a_i \right\|^2;$$

$\omega(\mathcal{Y})$ étant définie sur l'ensemble des indices $\mathcal{Y}$, $\mathcal{Y} \subset \mathcal{J}$. Vu cette inclusion, $\omega(\mathcal{Y})$ ne peut prendre qu'un nombre fini de valeurs. Il résulte de (5.41) que

$$\omega(\tilde{\mathcal{J}}(x_k)) \to 0.$$

Or, cela signifie que $\omega(\tilde{\mathcal{J}}(x_k)) = 0$ pour tous les k suffisamment grands car $\omega(\mathcal{Y})$ parcourt, nous venons de le dire, des valeurs en nombre fini.

Ainsi, pour k grands,

$$\omega(\tilde{\mathcal{J}}(x_k)) = 0. \quad (5.42)$$

Choisissons un k si élevé que $\alpha_k = 1$ et la condition (5.42) a lieu, ainsi que l'inclusion $\tilde{\mathcal{J}}(x_k) \subset \mathcal{J}_0(x_*)$. Comme $\alpha_k = 1$, on a $x_{k+1} = = x_k + p_k$. Etant donné $x_k \to x_*$, $p_k \to 0$, on admet que

$$f_i(x_{k+1}) \leqslant \frac{\varepsilon_0}{2} < 0, \ i \notin \mathcal{J}_0(x_*). \quad (5.43)$$

Revenons au problème auxiliaire (5.37). p_k vérifiant les contraintes (5.37) et $f_i(x)$ étant linéaires, on a

$$f_i(x_{k+1}) = (f_i'(x_k), p_k) + f_i(x_k) \leqslant 0 \tag{5.44}$$

pour $i \in \mathcal{J}_\delta(x_k)$ et, partant, pour $i \in \mathcal{J}_0(x_*)$ car $\mathcal{J}_0(x_*) \subset \mathcal{J}_\delta(x_k)$. Cela montre que x_{k+1} satisfait à toutes les contraintes du problème (5.2).

Montrons que x_{k+1} est solution de ce problème. En effet, (5.38) et la définition de l'ensemble $\tilde{\mathcal{J}}(x_k)$ entraînent

$$f_i(x_{k+1}) = 0, \quad i \in \tilde{\mathcal{J}}(x_k). \tag{5.45}$$

Mais l'égalité (5.42) signifie qu'il existe des nombres $u_0^i \geqslant 0$, $i \in \tilde{\mathcal{J}}(x_k)$, tels que

$$a_0 + \sum_{i \in \tilde{\mathcal{J}}(x_k)} u_0^i a_i = 0. \tag{5.46}$$

Posons $u_0^i = 0$, $i \notin \tilde{\mathcal{J}}(x_k)$; nous obtenons l'existence des nombres $u_0^i \geqslant 0$ vérifiant les conditions

$$a_0 + \sum_{i \in \mathcal{J}} u_0^i a_i = 0, \quad u_0^i f_i(x_{k+1}) = 0.$$

Mais les dernières relations sont (voir ch. I, § 3) les conditions nécessaires et suffisantes pour que le point x_{k+1} soit solution du problème de programmation linéaire.

Ainsi, l'algorithme donne une solution après un nombre fini de pas, c.q.f.d.

7. Evaluation locale de la convergence. On a montré au n° 6 que dans le cas linéaire l'algorithme proposé converge en un nombre fini de pas. Nous allons établir que s'agissant du cas non linéaire général et des conditions naturelles la convergence est en progression géométrique et même quadratique dans certaines circonstances favorables.

THÉORÈME 5.4. *Soit x_* solution du problème* (5.2). *Si*

a) *quel que soit $\delta > 0$ suffisamment petit, le problème auxiliaire* (5.4) *admet une solution*;

b) *les fonctions $f_i(x)$ sont deux fois continûment dérivables et les gradients $f_i'(x_*)$, $i \in \mathcal{J}_0(x_*)$, avec*

$$\mathcal{J}_0(x_*) = \{i : f_i(x_*) = 0, \ i \in \mathcal{J}\},$$

sont linéairement indépendants;

c) *le point x_* vérifie une condition nécessaire de minimum de la forme*

$$f_0'(x_*)+\sum_{i\in\mathcal{J}_0(x_*)} u_0^i f_i'(x_*)=0$$

et $u_0^i>0$, $i\in\mathcal{J}_0(x_*)$;

d) *on remplit une condition suffisante de minimum local, i.e.*

$$(p, L''(x_*, u_0)\,p)>0$$

pour tous les $p\neq 0$ *vérifiant la condition*

$$(p, f_i'(x_*))=0, \quad i\in\mathcal{J}_0(x_*),$$

où

$$L(x, u)=f_0(x)+\sum_{i\in\mathcal{J}_0(x_*)} u^i f_i(x),$$

et L'' *désigne la matrice des dérivées secondes* $L(x, u)$ *par rapport à* x, *il existe alors un voisinage* Ω *du point* x_*, $\delta_0>0$, *et un* $\alpha>0$ *tels que le processus*

$$x_{k+1}=x_k+\alpha p_k \tag{5.47}$$

converge vers x_* *à partir de toute approximation initiale* $x_0\in\Omega$ *et cette convergence est géométrique, i.e. il existe un nombre* $0\leqslant q<1$ *tel que* $\|x_*-x_k\|\leqslant Cq^k$ *pour tous les* k *suffisamment grands.*

Démonstration. L'idée de base est la suivante. Nous savons que le point x_* vérifie l'équation

$$p(x_*)=0.$$

(5.47) est un processus itératif simple destiné à la résolution de cette équation. La convergence s'évalue donc à l'aide du théorème d'Ostrowski qui sera énoncé plus bas et qui exige qu'on évalue les valeurs propres de la matrice des dérivées premières de $p(x)$ en x_*. Par conséquent, notre tâche consiste en premier lieu à calculer cette matrice et ses valeurs propres.

La démonstration du théorème se fera en plusieurs étapes.

Posons

$$\mathcal{J}_0(x_*)=\{i\in\mathcal{J}:\ f_i(x_*)=0\},$$
$$\varepsilon_0=\max_{i\notin\mathcal{J}_0(x_*)} f_i(x_*)<0.$$

Lemme 5.2. *Etant donné les conditions du théorème et l'inégalité* $\delta<-\frac{\varepsilon_0}{2}$, *il existe un voisinage du point* x_* *tel que* $\mathcal{J}_\delta(x)=\mathcal{J}_0(x_*)$ *et* $p(x)$ *y soit continûment dérivable par rapport à* x. *De plus, l'ensemble*

$$\tilde{\mathcal{J}}(x)=\{i\in\mathcal{J}_\delta(x):(f_i'(x), p(x))+f_i(x)=0\}$$

coïncide avec $\mathcal{J}_0(x_*)$.

DÉMONSTRATION. Vu la continuité de toutes les fonctions $f_i(x)$, il existe un voisinage de x_* tel que

$$-\delta/2 < f_i(x) < \delta/2, \quad i \in \mathcal{J}_0(x_*), \tag{5.48}$$

$$f_i(x) < \varepsilon_0/2, \quad i \notin \mathcal{J}_0(x_*). \tag{5.49}$$

Rappelons que

$$F(x) = \max\{0, \max_{i \in \mathcal{J}} f_i(x)\}$$

et $i \in \mathcal{J}_\delta(x)$ si $f_i(x) \geqslant F(x) - \delta$. Il découle de (5.48), (5.49) que

$$0 \leqslant F(x) \leqslant \delta/2, \tag{5.50}$$

et si $i \notin \mathcal{J}_0(x_*)$, alors

$$F(x) - \delta \geqslant -\delta > \varepsilon_0/2 > f_i(x),$$

i.e. $i \notin \mathcal{J}_\delta(x)$. D'autre part, pour $i \in \mathcal{J}_0(x_*)$, (5.50) entraîne $F(x) - \delta < -\delta/2$ et donc

$$f_i(x) > -\delta/2 > F(x) - \delta,$$

i.e. $i \in \mathcal{J}_\delta(x)$.

Ainsi, $\mathcal{J}_\delta(x) = \mathcal{J}_0(x_*)$ dans un voisinage de x_*.

Rappelons que si $p(x)$ est solution du problème (5.4), on a les conditions (5.6) qui se récrivent sous forme équivalente

$$p(x) + f_0'(x) + \sum_{i \in \tilde{\mathcal{J}}(x)} u^i(x) f_i'(x) = 0, \tag{5.51.1}$$

$$(f_i'(x), p(x)) + f_i(x) = 0, \quad i \in \tilde{\mathcal{J}}(x), \tag{5.51.2}$$

$$(f_i'(x), p(x)) + f_i(x) < 0, \quad i \notin \tilde{\mathcal{J}}(x), \ i \in \mathcal{J}_\delta(x), \tag{5.51.3}$$

avec $u^i(x) \geqslant 0$.

Introduisons plusieurs nouvelles notations. Soit $\mathcal{Y} \subset \mathcal{J}_\delta(x)$ $(= \mathcal{J}_0(x_*))$. Désignons par $f'_{\mathcal{Y}}(x)$ la matrice de lignes $f_i'(x)$, $i \in \mathcal{Y}$, par $f_{\mathcal{Y}}(x)$ le vecteur-colonne de composantes $f_i(x)$, $i \in \mathcal{Y}$, et par $u_{\mathcal{Y}}$ le vecteur-colonne de composantes u^i, $i \in \mathcal{Y}$. Les équations (5.51), (5.52) prennent alors la forme

$$\begin{gathered} p(x) + f_0'(x) + f'^*_{\mathcal{Y}}(x) u_{\mathcal{Y}}(x) = 0, \\ f'_{\mathcal{Y}}(x) p(x) + f_{\mathcal{Y}}(x) = 0, \quad \mathcal{Y} = \tilde{\mathcal{J}}(x). \end{gathered} \tag{5.52}$$

Ces relations peuvent s'interpréter comme système linéaire d'équations en $p(x)$ et $u_{\mathcal{Y}}(x)$. On voit sans peine que dans un certain voisinage de x_* le système (5.52) possède une solution unique donnée par les formules

$$\begin{gathered} u_{\mathcal{Y}}(x) = (f'_{\mathcal{Y}}(x) f'^*_{\mathcal{Y}}(x))^{-1} [f_{\mathcal{Y}}(x) - f'_{\mathcal{Y}}(x) f_0'(x)], \\ p(x) = -f_0'(x) - f'^*_{\mathcal{Y}}(x) u(x). \end{gathered} \tag{5.53}$$

Il en résulte la dépendance continue de $u_{\mathcal{Y}}(x)$ et $p(x)$ par rapport à x si l'ensemble $\mathcal{Y}$ est fixé.

Soit $x_k \to x_*$. Montrons que $\tilde{\mathcal{J}}(x_k) = \mathcal{J}_0(x_*)$ pour tous les k importants.

Admettons que notre hypothèse est fausse et qu'il existe des indices k aussi grands qu'on le veut et tels que $\tilde{\mathcal{J}}(x_k)$ est un sous-ensemble propre de $\mathcal{J}_0(x_*)$. Puisque $\tilde{\mathcal{J}}(x)$ distincts sont en nombre fini, on dit sans resteindre la généralité que la suite $x_k \to x_*$ choisie est telle que $\tilde{\mathcal{J}}(x_k) = \mathcal{Y}$, $\mathcal{Y} \subset \mathcal{J}_0(x_*)$.

Portons x_k au lieu de x dans (5.51) et passons à la limite ($p(x_k) \to \tilde{p}$, $u^i(x_k) \to \tilde{u}^i$, $i \in \mathcal{Y}$), il vient

$$\tilde{p} + f_0'(x_*) + \sum_{i \in \mathcal{Y}} \tilde{u}^i f_i'(x_*) = 0,$$

$$(f_i'(x_*), \tilde{p}) + f_i(x_*) = 0, \quad i \in \mathcal{Y},$$

$$(f_i'(x_*), \tilde{p}) + f_i(x_*) \leqslant 0, \quad i \notin \mathcal{Y}, \ i \in \mathcal{J}_0(x_*) = \mathcal{J}_\delta(x_*),$$

et $\tilde{u}^i \geqslant 0$ car $u^i(x_k) \geqslant 0$. Mais les dernières relations prouvent que $\tilde{p}$ est solution en x_* du problème auxiliaire (5.4), i.e. $\tilde{p} = p(x_*)$. Or, x_* est solution de (5.2) et donc $p(x_*) = 0$. Ainsi

$$f_0'(x_*) + \sum_{i \in \mathcal{Y}} \tilde{u}^i f_i'(x_*) = 0.$$

Utilisons la condition c) du théorème. La dernière relation donne alors

$$\sum_{i \in \mathcal{Y}} (u_0' - \tilde{u}^i) f_i'(x_*) + \sum_{i \in \mathcal{J}_0(x_0) \setminus \mathcal{Y}} u_0^i f_i'(x_*) = 0,$$

ce qui contredit la condition b). Ainsi, dans un voisinage du point x_* l'ensemble $\tilde{\mathcal{J}}(x)$ coïncide avec $\mathcal{J}_0(x_*)$. Cette constance de $\tilde{\mathcal{J}}(x)$ et (5.53) entraîne de suite la dérivabilité continue de $u_{\mathcal{Y}}(x)$ et $p(x)$, $\mathcal{Y} = \mathcal{J}_0(x_*)$ par rapport à x parce qu'en vertu de b) $f_i(x)$ sont deux fois continûment dérivables.

Remarque. Ainsi, dans un voisinage restreint de x_*, $p(x)$ et $u_{\mathcal{Y}}(x)$ est solution du système (5.52) pour l'ensemble $\mathcal{Y} = \mathcal{J}_0(x_*)$ constant (ce qui permet d'omettre l'indice $\mathcal{Y}$ de $u_{\mathcal{Y}}(x)$).

Lemme 5.3. *La matrice $p'(x)$ des dérivées du vecteur $p(x)$, i.e. la matrice d'éléments $\partial p^i(x)/\partial x^j$, $i, j = 1, \ldots, n$, avec $p^i(x)$ la i-ième composante de $p(x)$, s'écrit en x_*:*

$$p'(x_*) = -[P + (I - P) L''(x_*, u_0)],$$

où $P = f'^{*}_{\mathcal{J}_0(x_*)}(x_*)\,(f'_{\mathcal{J}_0(x_*)}(x_*)\,f'^{*}_{\mathcal{J}_0(x_*)}(x_*))^{-1}\,f'_{\mathcal{J}_0(x_*)}(x_*)$,
et $u_0 = u(x_*)$.

Démonstration. On obtient par dérivation immédiate de la première formule (5.52):

$$p'(x_*) = -L''(x_*, u_0) - \sum_{j \in \mathcal{J}_0(x_*)} f'_j(x_*)\,(u^{j\prime}(x_*))^*, \tag{5.54}$$

avec

$$u^{j\prime}(x) = \begin{pmatrix} \dfrac{\partial u^j(x)}{\partial x^1} \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{\partial u^j(x)}{\partial x^n} \end{pmatrix}.$$

La formule (5.51.2) donne par dérivation ($p(x_*) = 0$)

$$f'^{*}_i(x_*)\,p'(x_*) + f'^{*}_i(x_*) = 0, \quad i \in \mathcal{J}_0(x_*). \tag{5.55}$$

Indiquons que l'opérateur P défini dans le lemme est celui de projection sur un sous-espace engendré par les vecteurs $f'_i(x_*)$, $i \in \mathcal{J}_0(x_*)$. Cela résulte en effet (voir également ch. III, § 1, n° 1) des relations faciles à vérifier:

1) $Pf'^{*}_{\mathcal{J}_0(x_*)}(x_*) = f'^{*}_{\mathcal{J}_0(x_*)}(x_*)$ ou $Pf'_i(x_*) = f'_i(x_*)$, $i \in \mathcal{J}_0(x_*)$;
2) $P^* = P$, $P^2 = P$;
3) $(I - P)P = 0$.

En récrivant (5.55) comme

$$f'_{\mathcal{J}_0(x_*)}(x_*)\,p'(x_*) + f'_{\mathcal{J}_0(x_*)}(x_*) = 0,$$

on obtient à présent, compte tenu de l'expression de P,

$$Pp'(x_*) = -P. \tag{5.56}$$

La relation 1) entraîne ensuite $(I - P)\,f'_i(x_*) = 0$. Appliquons $(I - P)$ aux deux membres de (5.54), il vient

$$(I - P)\,p'(x_*) = -(I - P)\,L''(x_*, u_0). \tag{5.57}$$

L'addition de (5.56) et (5.57) fournit la formule cherchée pour $p'(x_*)$.

Lemme 5.4. *Les valeurs propres* γ_j *de la matrice* $p'(x_*)$ *peuvent être individualisées comme suit:* $\gamma_j = -1$ *pour* $j = 1, 2, \ldots, m$, *avec* $m \leqslant n$ *le nombre d'indices de l'ensemble* $\mathcal{J}_0(x_*)$; $\gamma_j = -\lambda_{j-m}$, $j = m + 1, \ldots, n$, *avec* λ_j, $j = 1, \ldots, n - m$, *les valeurs propres de la matrice* $(I - P)\,L''(x_*, u_0)\,(I - P)$, *et* $\lambda_j > 0$, $j = 1, \ldots, n - m$.

DÉMONSTRATION. Soient σ une valeur propre et y un vecteur propre de la matrice $p'(x_*)$. Selon le lemme 5.3, on a alors

$$-Py - (I - P) L''(x_*, u_0) y = \sigma y = \sigma P y + \sigma (I - P) y.$$

Utilisons la relation $P(I - P) = 0$ et multiplions la dernière égalité par P, puis par $I - P$, il vient

$$-Py = \sigma P y, \tag{5.58}$$

$$-(I - P) L''(x_*, u_0) y = \sigma (I - P) y. \tag{5.59}$$

Deux cas peuvent se présenter:

1) $Py \neq 0$. Il résulte alors de (5.58) que $\sigma = -1$.

2) $Py = 0$. Dans ce cas $(I - P) y = y$ et (5.59) se récrit

$$(I - P) L''(x_*, u_0) (I - P) y = -\sigma y, \tag{5.60}$$

i.e. σ est valeur propre de la matrice $(I - P) L''(I - P)$. Cette dernière est symétrique vu que $P = P^*$ et $L'' = (L'')^*$ en tant que matrice des dérivées secondes de la fonction L. De plus, la matrice considérée est définie non négative. En effet, quel que soit w,

$$(w, (I - P) L''(I - P) w) = (z, L''z),$$

où

$$z = (I - P) w.$$

Mais $f'_{\mathcal{J}_0(x_*)}(x_*) z = f'_{\mathcal{J}_0(x_*)}(x_*) (I - P) w = 0$ et donc $(z, L''z) \geqslant 0$ en vertu de la condition d) du théorème 5.4, l'égalité n'ayant lieu que si $z = (I - P) w = 0$. La symétrie de la matrice $(I - P) L''(I - P)$ entraîne le caractère réel de ses valeurs et vecteurs propres. Comme $y \neq 0$ et $y = Py + (I - P) y$, il découle de $Py = 0$ que $(I - P) y \neq 0$ et (5.60) fait donc que

$$-\sigma (y, y) = (y, (I - P) L''(I - P) y) = (y, L''y) > 0.$$

Ainsi, $-\sigma \neq 0$ et donc $\sigma = -\lambda_j$, où $\lambda_j > 0$ est valeur propre de $(I - P) L''(I - P)$.

Nous avons démontré que les valeurs propres de la matrice $p'(x_*)$ sont réelles et valent soit -1, soit $-\lambda_j$, $\lambda_j > 0$. Il nous reste à établir le nombre de valeurs propres égales à -1.

Par suite de la condition

$$P f'_i(x_*) = f'_i(x_*), \qquad i \in \mathcal{J}_0(x_*),$$

l'opérateur $(I - P)$ possède m vecteurs propres $f'_i(x_*)$ associés à une valeur propre nulle. Aussi la matrice $(I - P) L''(I - P)$ a également m valeurs propres nulles. On a vu d'autre part que $p'(x)$ a toutes ses n valeurs propres non nulles, dont chacune est soit -1, soit une valeur propre non nulle de la matrice $(I - P) \times \times L''(I - P)$. On conçoit que cela n'a lieu que dans le cas de la validité de l'affirmation du lemme 5.4.

Nous disposons maintenant de tous les éléments nécessaires pour achever la démonstration du théorème 5.4. Il s'ensuit du théorème d'Ostrowski [91] que si x_* est solution de l'équation $p(x) = 0$ et les valeurs propres de la matrice $I + \alpha p'(x_*)$ sont de module inférieur à l'unité, la méthode itérative simple $x_{k+1} = x_k + \alpha p(x_k)$ converge à partir de tout point d'un voisinage de x_* et on a l'estimation : pour chaque $\varepsilon > 0$ il existe un nombre $C(\varepsilon)$ tel que $\| x_* - x_k \| \leqslant C(\varepsilon)(q_0 + \varepsilon)^k$, avec q_0 le plus grand module des valeurs propres de $I + \alpha p'(x_*)$.

Considérons maintenant les valeurs propres de la matrice $I + \alpha p'(x_*)$. Elles valent soit $1 - \alpha$, soit $1 - \alpha\lambda_j$. Choisissons un α tel que l'on ait toutes les inégalités

$$1 - \alpha > -1, \; 1 - \alpha\lambda_j > -1, \; j = 1, \ldots, n - m,$$

i.e. $0 < \alpha < \min\{2, 2/\lambda_0\}$, où $\lambda_0 = \max \lambda_j$, $j = 1, \ldots, n - m$. Toutes les valeurs propres de $I + \alpha p'(x_*)$ sont alors de module inférieur à 1, d'où, en se référant au résultat mentionné d'Ostrowski, le théorème 5.4.

THÉOREME 5.5. *Si, dans les conditions du théoreme* 5.4, *m (le nombre d'indices de l'ensemble $\mathcal{J}_0(x_*)$) est égal à n (la dimension de l'espace), la convergence du processus* (5.47) *à partir d'un voisinage du point x_* pour $\alpha = 1$ est quadratique.*

DÉMONSTRATION. Le lemme 5.4 entraîne pour le cas considéré que toutes les valeurs propres de la matrice $p'(x_*)$ sont égales à -1, et donc celles de $I + \alpha p'(x_*)$ valent $1 - \alpha$. Si $\alpha = 1$, toutes les valeurs propres valent 0 et $q_0 = 0$. Il résulte donc du théorème d'Ostrowski que $\| x_* - x_k \| \leqslant C(\varepsilon)\,\varepsilon^k$, ce qui signifie qu'on a une convergence plus rapide qu'en progression géométrique. En fait, le processus (5.47) devient alors la méthode de Newton pour le système d'équations $f_i(x) = 0$, $i \in \mathcal{J}_0(x_*)$, dont la convergence est quadratique (propriété établie sur laquelle nous reviendrons au § 6).

REMARQUE. Tous les raisonnements du numéro ont porté sur un problème avec des contraintes inégalité. Mais les résultats se généralisent évidemment au cas comportant des contraintes de type égalité.

§ 6. Méthode de linéarisation (résolution de systèmes d'égalités et d'inégalités et recherche du minimax)

Dans ce paragraphe nous appliquerons la méthode de linéarisation à deux problèmes étroitement liés à des problèmes usuels de programmation mathématique. On constate en l'occurrence la possibilité d'algorithmes efficaces présentant une bonne convergence.

1. Systèmes d'égalités et d'inégalités. Etant donné deux familles finies d'indices $\mathcal{J}^-$ et $\mathcal{J}^0$ et les fonctions $f_i(x)$, $x \in E^n$, on demande

la solution du système

$$f_i(x) \leqslant 0, \quad i \in \mathcal{J}^-, \quad f_i(x) = 0, \quad i \in \mathcal{J}^0. \tag{6.1}$$

Les fonctions $f_i(x)$ seront supposées posséder des gradients $f_i'(x)$ continus qui vérifient la condition de Lipschitz de constante L:

$$\| f_i'(x_1) - f_i'(x_2) \| \leqslant L \| x_1 - x_2 \|.$$

La norme des vecteurs est partout euclidienne.

Introduisons les notations

$$F(x) = \max(\max_{i \in \mathcal{J}^-} f_i(x), \max_{i \in \mathcal{J}^0} |f_i(x)|),$$

$$\mathcal{J}_\delta^-(x) = \{i : i \in \mathcal{J}^-, f_i(x) \geqslant F(x) - \delta\},$$

$$\mathcal{J}_\delta^0(x) = \{i : i \in \mathcal{J}^0, |f_i(x)| \geqslant F(x) - \delta\}.$$

Soit x_0 un point initial choisi et supposons que pour tous les x satisfaisant à l'inégalité $F(x) \leqslant F(x_0)$ les gradients $f_i'(x)$ sont bornés en norme par une constante K.

Hypothese fondamentale. Il existe des nombres $\delta > 0$ et $C > 0$ tels que, quel que soit x pour lequel $F(x) > 0$, $F(x) \leqslant F(x_0)$, le système

$$\begin{aligned} &(f_i'(x), p) + f_i(x) \leqslant 0, \quad i \in \mathcal{J}_\delta^-(x), \\ &(f_i'(x), p) + f_i(x) = 0, \quad i \in \mathcal{J}_\delta^0(x), \end{aligned} \tag{6.2}$$

soit résoluble par rapport à p. Soit $p(x)$ solution de (6.2) munie de la norme minimale. On a alors pour des x tels que $F(x) > 0$:

$$\| p(x) \| \leqslant CF(x). \tag{6.3}$$

L'inégalité (6.3) caractérise dans une certaine mesure la résolubilité régulière du système (6.2). En particulier, si celui-ci devient n équations à n inconnues, la condition (6.3) équivaut à supposer la régularité de la matrice du système correspondant. On montrera dans la suite que (6.3) a lieu si les gradients $f_i'(x)$, $i \in \mathcal{J}_\delta^-(x) \cup \mathcal{J}_\delta^0(x)$, sont linéairement indépendants pour tous les x, $F(x) > 0$.

Construisons l'algorithme. Les approximations successives sont générées par la formule

$$x_{k+1} = x_k + \alpha_k p_k, \quad p_k \equiv p(x_k), \tag{6.4}$$

avec le paramètre α_k choisi par bipartition de l'unité jusqu'à ce qu'on ait pour la première fois

$$F(x_k + \alpha_k p_k) \leqslant (1 - \varepsilon\alpha_k) F(x_k), \tag{6.5}$$

où ε est un nombre quelconque choisi d'avance, $0 < \varepsilon < 1$. Il est clair que la formule (6.4) s'applique quand $F(x) > 0$. Dans le cas contraire, le processus s'arrête et x_k se trouve être solution de (6.1).

2. Convergence de l'algorithme. La procédure de calcul proposée est caractérisée par le

THÉORÈME 6.1. *Sous toutes les hypothèses du n° 1 la suite x_k, $k = 0, 1, \ldots$, générée par l'algorithme selon la formule* (6.4) *converge vers $\bar{x}$, solution du système* (6.1), *et*

a) $\alpha_k = 1$ *pour des k suffisamment grands*;

b) $F(x_{k+1}) \leqslant LC^2F^2(x_k)$ *pour des k suffisamment grands*;

c) *quel que soit q, $0 < q < 1$, il existe un numéro $k(q)$ tel que*

$$\|\bar{x} - x_k\| \leqslant \frac{q^{2^{k-k(q)}}}{LC(1-q)} \tag{6.6}$$

pour tous les $k \geqslant k(q)$.

DÉMONSTRATION. La démonstration est immédiate pour $F(x_k) \leqslant 0$ à un pas. Aussi nous supposerons $F(x_k) > 0$ pour tous les k.

Commençons par montrer qu'il est toujours possible de prendre un α_k en se basant sur la condition (6.5). Quand $i \in \mathcal{J}_\delta^-(x_k)$ on a moyennant la formule de Taylor

$$f_i(x_k + \alpha p_k) = f_i(x_k) + \alpha(f_i'(x_k + \theta_i \alpha p_k), p_k) =$$
$$= f_i(x_k) + \alpha(f_i'(x_k), p_k) + \alpha(f_i'(x_k + \theta_i \alpha p_k) - f_i'(x_k), p_k),$$

où $0 \leqslant \theta_i \leqslant 1$. Du moment que p_k vérifie (6.2),

$$(f_i'(x_k), p_k) \leqslant -f_i(x_k).$$

Ensuite,

$$(f_i'(x_k + \theta_i \alpha p_k) - f_i'(x_k), p_k) \leqslant \|p_k\| \, \|f_i'(x_k + \theta_i \alpha p_k) -$$
$$- f_i'(x_k)\| \leqslant \|p_k\| \, \|\theta_i \alpha p_k\| L \leqslant \alpha L \|p_k\|^2.$$

On obtient donc à l'aide de (6.3)

$$f_i(x_k + \alpha p_k) \leqslant f_i(x_k) - \alpha f_i(x_k) + \alpha^2 L \|p_k\|^2 \leqslant$$
$$\leqslant (1-\alpha) F(x_k) + \alpha^2 L c^2 F^2(x_k). \tag{6.7}$$

Pour $i \in \mathcal{J}^-$, $i \notin \mathcal{J}_\delta^-(x_k)$, $f_i(x_k) < F(x_k) - \delta$ et donc

$$f_i(x_k + \alpha p_k) = f_i(x_k) + \alpha(f_i'(x_k + \theta_i \alpha p_k), p_k) \leqslant$$
$$\leqslant F(x_k) - \delta + \alpha K \|p_k\| \leqslant F(x_k) - \delta + \alpha K C F(x_k). \tag{6.8}$$

De même pour $i \in \mathcal{J}_\delta^0(x_k)$ on a

$$|f_i(x_k + \alpha p_k)| \leqslant (1-\alpha) F(x_k) + \alpha^2 L C^2 F^2(x_k) \tag{6.9}$$

et pour $i \notin \mathcal{J}_\delta^0(x_k)$

$$|f_i(x_k + \alpha p_k)| \leqslant F(x_k) - \delta + \alpha C K F(x_k). \tag{6.10}$$

Notons que

$$(1-\alpha) F(x_k) \geqslant F(x_k) - \delta + \alpha C K F(x_k),$$

si $\alpha \leqslant \alpha_k^1$, où

$$\alpha_k^1 = \frac{\delta}{(1+CK)\,F(x_k)}.$$

Les relations (6.7)-(6.10) entraînent donc pour $\alpha \leqslant \alpha_k^1$

$$F(x_k + \alpha p_k) \leqslant (1-\alpha)\,F(x_k) + \alpha^2 L C^2 F^2(x_k)$$

ou

$$F(x_k + \alpha p_k) \leqslant F(x_k) - \alpha F(x_k)\,[1 - \alpha L C^2 F(x_k)]. \tag{6.11}$$

Et si $\alpha \leqslant \alpha_k^2$ avec

$$\alpha_k^2 = \frac{1-\varepsilon}{LC^2 F(x_k)},$$

alors $1 - \alpha L C^2 F(x_k) \geqslant \varepsilon$, et (6.11) se récrit donc:

$$\begin{gathered} F(x_k + \alpha p_k) \leqslant F(x_k) - \alpha\varepsilon F(x_k), \\ \alpha \leqslant \min\{\alpha_k^1, \alpha_k^2\}. \end{gathered} \tag{6.12}$$

On conçoit maintenant qu'en divisant α par deux à commencer par $\alpha = 1$ on a l'inégalité (6.12) après un nombre fini d'essais et α_k choisi vérifie

$$\alpha_k \geqslant \min\left\{1, \frac{1}{2}\alpha_k^1, \frac{1}{2}\alpha_k^2\right\}. \tag{6.13}$$

Cela démontre du même coup la possibilité de choisir α_k à partir de la condition (6.5) et le fait que ce choix s'effectue en un nombre fini d'opérations.

Montrons que $F(x_k) \to 0$. En effet, (6.5) implique la décroissance monotone de $F(x_k)$. On déduit donc des formules pour α_k^1 et α_k^2 que ces quantités croissent avec k. La formule (6.13) permet donc de conclure à $\alpha_k \geqslant \bar{\alpha} > 0$, de sorte que

$$F(x_{k+1}) \leqslant (1 - \varepsilon\alpha_k)\,F(x_k) \leqslant (1 - \varepsilon\bar{\alpha})\,F(x_k).$$

Aussi $F(x_k) \leqslant (1 - \varepsilon\bar{\alpha})^k F(x_0)$, d'où $F(x_k) \to 0$. Mais on a alors $\alpha_k^1 \to +\infty$, $\alpha_k^2 \to +\infty$, comme cela résulte de suite des formules donnant ces quantités. On établit donc moyennant (6.13) que $\alpha_k = 1$ pour des k suffisamment grands. Or, pour de tels k on déduit de (6.11), avec $\alpha = 1$,

$$F(x_{k+1}) \leqslant LC^2 F^2(x_k). \tag{6.14}$$

Ainsi, nous venons de démontrer les affirmations a) et b) du théorème.

Nous sommes maintenant autorisés à dire qu'il existe un k_0 tel que $\alpha_k = 1$ pour $k \geqslant k_0$ et qu'on ait (6.14). C'est pourquoi, on a en vertu de (6.3)

$$\| x_{k+1} - x_k \| = \| p_k \| \leqslant CF(x_k).$$

BIBLIOTHÈQUE

Posons $v_k = LC^2F(x_k)$. Alors $v_k \to 0$ et (voir (6.14)) $v_{k+1} \leqslant v_k^2$. Soit q tel que $0 < q < 1$. On peut alors trouver un $k(q)$ tel que $v_k < q$ pour $k \geqslant k(q)$. Donc $v_{k+1} \leqslant qv_k$, $k \geqslant k(q)$. D'où

$$v_k \leqslant v_{k(q)}^{2^{k-k(q)}} \leqslant q^{2^{k-k(q)}}, \quad v_m \leqslant q^{m-k} v_k, \quad m \geqslant k \geqslant k(q).$$

Cela permet d'obtenir l'estimation

$$\|x_m - x_k\| \leqslant \sum_{j=k}^{m-1} \|x_{j+1} - x_j\| \leqslant \frac{1}{LC} \sum_{j=k}^{m-1} v_j \leqslant$$

$$\leqslant \frac{v_k}{LC} \sum_{j=0}^{m-1-k} q^j \leqslant \frac{v_k}{LC(1-q)} \leqslant \frac{q^{2^{k-k(q)}}}{LC(1-q)}. \tag{6.15}$$

Il en résulte (d'après le critère connu de Cauchy) la convergence de la suite x_k vers un point $\bar{x}$. Comme $F(x_k) \to 0$, on a $F(\bar{x}) = 0$, i.e. $\bar{x}$ est solution du système (6.1). Par passage à la limite pour $m \to \infty$ dans (6.15), on obtient de plus

$$\|\bar{x} - x_k\| \leqslant \frac{q^{2^{k-k(q)}}}{LC(1-q)},$$

c.q.f.d.

3. Remarques.

Remarque 1. Soit à résoudre le système de n équations $f_i(x) = 0$, $i = 1, \ldots, n$, où $x \in E^n$. Alors

$$f_i(x) \geqslant F(x) - \delta, \quad i = 1, \ldots, n,$$
$$F(x) = \max_{1 \leqslant i \leqslant n} |f_i(x)|$$

pour δ arbitraire à condition que x soit suffisamment voisin de la solution $\bar{x}$. Aussi $\mathcal{J}_\delta^0(x) = \{1, 2, \ldots, n\}$ et le système (6.2) prend la forme

$$(f_i'(x), p) + f_i(x) = 0, \quad i = 1, \ldots, n. \tag{6.16}$$

C'est pourquoi la technique proposée coïncide tout simplement avec la méthode de Newton dans laquelle on itère suivant la formule $x_{k+1} = x_k + p(x_k)$ avec $p(x)$ solution du système (6.16). La condition de convergence de ladite méthode est la régularité en $\bar{x}$ de la matrice $f'(\bar{x})$, où $f'(x)$ est une matrice $n \times n$ dont les lignes sont formées par $f_i'(x)$. Dans ce cas $p(x) = -(f'(x))^{-1} f(x)$, $f(x)$ étant un vecteur-colonne de composantes $f_i(x)$. Cette dernière formule entraîne

$$\|p(x)\| \leqslant \|(f'(x))^{-1}\| \, \|f(x)\| \leqslant C_0 \|(f'(x))^{-1}\| \, F(x).$$

où C_0 est une constante. Cette inégalité montre que (6.3) a lieu dans un voisinage du point $\bar{x}$.

Le théorème démontré implique donc la convergence locale de la méthode classique de Newton pour un système de n équations à n inconnues.

Remarque 2. Si l'on a une équation $f(x) = 0$ à n inconnues, le système (6.2) s'écrit comme

$$(f'(x), p) + f(x) = 0, \tag{6.17}$$

et on demande de trouver la solution de cette équation qui soit munie de la norme minimale, i.e. trouver le minimum de $\| p \|^2$ dans les contraintes (6.17). La règle des multiplicateurs de Lagrange aidant on obtient facilement qu'alors

$$p(x) = -\frac{f(x)}{\| f'(x) \|^2} f'(x),$$

d'où

$$\| p(x) \| = \frac{1}{\| f'(x) \|} | f(x) |.$$

Il est clair que la formule (6.3) est vraie si $\| f'(x) \| \geqslant \gamma$ quels que soient x.

Remarque. 3. La recherche du vecteur $p(x)$ à chaque pas est liée à celle du minimum de $\| p \|^2$ sous les contraintes (6.2). C'est un problème de programmation quadratique. On le résout en s'inspirant des indications du § 5 relatives au problème auxiliaire de programmation quadratique intervenant dans la méthode de linéarisation.

4. Conditions suffisantes de convergence. L'hypothèse fondamentale (6.3) qui garantit la convergence de l'algorithme se prête mal à une vérification. Les conditions données dans ce numéro sont testées plus facilement. En particulier, la convergence est garantie dans le cas convexe si le domaine défini par (6.1) possède un point intérieur.

Supposons que le système ne renferme que des contraintes inégalité, i.e. il s'écrit

$$f_i(x) \leqslant 0, \quad i \in \mathcal{J}^-. \tag{6.18}$$

Le système auxiliaire (6.2) prend alors la forme

$$(f_i'(x), p) + f_i(x) \leqslant 0, \quad i \in \mathcal{J}_{\bar{\delta}}(x). \tag{6.19}$$

Il est clair qu'il admet une solution pour $F(x) > 0$ à condition que

$$(f_i'(x), p) + F(x) \leqslant 0, \quad i \in \mathcal{J}_{\bar{\delta}}(x), \tag{6.20}$$

soit résoluble.

LEMME 6.1. *Si* $F(x) > 0$, *le système* (6.20) *possède une solution si et seulement si*

$$L_\delta(x) = \min_{\lambda_i \geqslant 0} \Big\| \sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i f_i'(x) \Big\| > 0,$$

où l'on minimise par rapport à tous les $\lambda_i \geqslant 0$ *tels que*

$$\sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i = 1.$$

Ceci étant, la solution $\bar{p}(x)$ *de norme minimale du système* (6.20) *satisfait à l'égalité*

$$\|\bar{p}(x)\| = \frac{1}{L_\delta(x)} F(x).$$

DÉMONSTRATION. Soit $\lambda_i \geqslant 0$ tels que leur somme en $i \in \mathcal{J}_\delta^-(x)$ soit égale à l'unité. Si p est une solution de (6.20), on a

$$-\sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i (f_i'(x), p) \geqslant F(x),$$

ou

$$\Big(-\sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i f_i'(x), p\Big) \geqslant F(x).$$

Utilisons l'inégalité $(x, y) \leqslant \|x\| \, \|y\|$, il vient

$$\Big\| \sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i f_i'(x) \Big\| \, \|p\| \geqslant F(x).$$

Mais la dernière inégalité est juste pour tout choix indiqué de λ_i, donc

$$L_\delta(x) \, \|p\| \geqslant F(x),$$

i.e. $L_\delta(x) > 0$ et

$$\|p\| \geqslant \frac{F(x)}{L_\delta(x)}. \tag{6.21}$$

Nous avons donc établi la nécessité.

Admettons que $L_\delta(x) > 0$.

Considérons le problème : trouver le minimum de ρ sous les contraintes

$$(f_i'(x), p) + F(x) - \rho \leqslant 0, \quad i \in \mathcal{J}_\delta^-(x), \tag{6.22}$$

$$\|p\| \leqslant r_0, \quad r_0 = \frac{F(x)}{L_\delta(x)} > 0.$$

C'est un problème de programmation convexe pour lequel sont immédiates toutes les conditions du théorème de Kuhn et Tucker, en particuler la condition de Slater. Soit p_0, ρ_0 une solution. Appli-

quons le théorème de Kuhn et Tucker. Il existe alors des $\lambda_i \geqslant 0$ tels que

$$\rho_0 + \sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i ((f_i'(x), p_0) + F(x) - \rho_0) \leqslant$$
$$\leqslant \rho + \sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i ((f_i'(x), p) + F(x) - \rho) \tag{6.23}$$

quels que soient p, $\| p \| \leqslant r_0$, et ρ. De plus

$$\lambda_i ((f_i'(x), p_0) + F(x) - \rho_0) = 0, \quad i \in \mathcal{J}_\delta^-(x). \tag{6.24}$$

ρ étant quelconque (6.23) entraîne de suite

$$\sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i = 1,$$

et (6.23) se récrit en vertu de (6.24) :

$$\rho_0 \leqslant \Big(\sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i f_i'(x), p \Big) + F(x).$$

Minimisons par rapport à p $\| p \| \leqslant r_0$ dans le second membre de la dernière inégalité, il vient

$$\rho_0 \leqslant -r_0 \Big\| \sum_{i \in \mathcal{J}_\delta^-(x)} \lambda_i f_i'(x) \Big\| + F(x) \leqslant -r_0 L_\delta(x) + F(x) = 0.$$

Ainsi, $\rho_0 \leqslant 0$, i.e. le vecteur p_0 vérifie le système (voir (6.22))

$$(f_i'(x), p_0) + F(x) \leqslant \rho_0 \leqslant 0, \quad i \in \mathcal{J}_\delta^-(x),$$

et $\| p_0 \| \leqslant r_0 = F(x)/L_\delta(x)$. Mais il découle de (6.21)

$$\| p_0 \| \geqslant \frac{F(x)}{L_\delta(x)}.$$

Donc

$$\| p_0 \| = \frac{F(x)}{L_\delta(x)}$$

et le vecteur p_0 est solution de (6.20). De plus, (6.21) montre la minimalité de sa norme. Ainsi, $p_0 = \bar{p}(x)$, ce qui démontre le lemme.

THÉORÈME 6.2. *Soit, sous toutes les hypothèses du* n° 1 *sauf l'hypothèse fondamentale,* $L_\delta(x) \geqslant \gamma > 0$ *pour tous les* x *tels que* $0 < F(x) \leqslant F(x_k)$. *L'hypothèse fondamentale est alors satisfaite elle aussi, ainsi que toutes les conclusions du théorème* 6.1 *pour le problème* (6.18).

DÉMONSTRATION. Puisque toute solution du système (6.20) est encore celle de (6.19), on a

$$\| p(x) \| \leqslant \| \bar{p}(x) \|.$$

En vertu du lemme 6.1,

$$\| p(x) \| \leqslant \frac{F(x)}{L_\delta(x)} \leqslant \frac{1}{\gamma} F(x);$$

ce qui montre qu'on est dans les conditions du théorème 6.1.

Notons que la condition $L_\delta(x) \geqslant \gamma > 0$ est suffisamment naturelle car elle exige l'indépendance linéaire des vecteurs $f_i'(x)$, $i \in \mathcal{J}_\delta^-(x)$.

THÉORÈME 6.3. *Si $f_i(x)$ du problème* (6.18) *sont des fonctions convexes continûment dérivables, le domaine défini par l'inégalité $F(x) \leqslant F(x_0)$ est compact, les gradients $f_i'(x)$ vérifient dans ce domaine la condition de Lipschitz et il existe un point $\bar{x}$ tel que $F(\bar{x}) = \gamma < 0$, toutes les conditions du théorème* 6.1 *sont alors satisfaites pour $\delta < -\gamma$.*

DÉMONSTRATION. Etant donné la convexité des $f_i(x)$,

$$f_i(\bar{x}) \geqslant f_i(x) + (f_i'(x), \bar{x} - x), \qquad i \in \mathcal{J}^-.$$

Quand $i \in \mathcal{J}_\delta^-(x)$, on a pour $p = \bar{x} - x$

$$f_i(\bar{x}) + \delta \geqslant f_i(x) + \delta + (f_i'(x), p).$$

Mais $f_i(\bar{x}) + \delta \leqslant F(\bar{x}) + \delta = \gamma + \delta < 0$ et $f_i(x) + \delta \geqslant F(x)$, $i \in \mathcal{J}_\delta^-(x)$. Donc

$$0 \geqslant \gamma + \delta \geqslant F(x) + (f_i'(x), p), \quad i \in \mathcal{J}_\delta^-(x).$$

Posons $\gamma + \delta = -\varepsilon$, il vient

$$(f_i'(x), p) + (F(x) + \varepsilon) \leqslant 0, \qquad i \in \mathcal{J}_\delta^-(x).$$

Or, cela signifie en vertu du lemme 6.1 que pour tous les x tels que $F(x) \leqslant F(x_0)$, $F(x) + \varepsilon \geqslant 0$, le système (6.20) possède une solution et $L_\delta(x) > 0$. Faisant jouer la compacité du domaine $F(x) \leqslant F(x_0)$ et la continuité on s'assure aisément que $L_\delta(x) \geqslant \gamma > 0$ pour tous les x tels que $0 \leqslant F(x) \leqslant F(x_0)$.

Toutes les conditions du théorème 6.2 se trouvent donc remplies, ce qui achève la démonstration du théorème 6.3.

5. Résolution d'un problème de minimax. Soient données les fonctions $f_i(x)$, $i = 1, \ldots, m$. Formons la fonction

$$F(x) = \max_{1 \leqslant i \leqslant m} f_i(x). \tag{6.25}$$

Le problème consiste à trouver un point $x \in E^n$ qui réalise le minimum de $F(x)$.

On voit que ce problème se ramène par introduction d'une variable supplémentaire x^{n+1} à la minimisation de $f_0(x, x^{n+1}) = x^{n+1}$

sous les contraintes

$$f_i(x) - x^{n+1} \leqslant 0, \quad i = 1, \ldots, m.$$

On recourt donc aux méthodes des paragraphes précédents, notamment à la méthode de linéarisation. Notons qu'on agit de même pour le problème de minimisation de $F(x)$ si x varie dans un domaine Ω défini par un système d'égalités ou d'inégalités.

Dans ce numéro nous nous occuperons d'une méthode de minimisation de $F(x)$ quand $x \in E^n$. Il s'agit de la méthode de linéarisation légèrement modifiée.

Faisons correspondre à tout point x le problème auxiliaire

$$\min\left(\beta + \frac{1}{2}\|p\|^2\right), \tag{6.26}$$

$$(f_i'(x), p) + f_i(x) - \beta \leqslant 0, \quad i \in \mathcal{J}_\delta(x),$$

où $\delta > 0$ et

$$\mathcal{J}_\delta(x) = \{i\colon 1 \leqslant i \leqslant m,\ f_i(x) \geqslant F(x) - \delta\}.$$

Notons que (6.26) est un problème de programmation convexe qui vérifie la condition de Slater : à condition de choisir un β suffisamment grand, les contraintes (6.26) sont satisfaites strictement. Appliquons directement le théorème de Kuhn et Tucker sous forme différentielle, il vient que $p(x)$ et $\beta(x)$ sont solution du problème (6.26) si et seulement si il existe des $u^i \geqslant 0$, $i \in \mathcal{J}_\delta(x)$ tels que

$$\sum_{i \in \mathcal{J}_\delta(x)} u^i = 1,$$

$$p(x) + \sum_{i \in \mathcal{J}_\delta(x)} u^i f_i'(x) = 0, \tag{6.27}$$

$$u^i((f_i'(x), p(x)) + f_i(x) - \beta(x)) = 0, \quad i \in \mathcal{J}_\delta(x).$$

Ensuite, le point $p = 0$, $\beta = F(x)$, vérifie évidemment les contraintes (6.26). Donc

$$\beta(x) + \frac{1}{2}\|p(x)\|^2 \leqslant F(x). \tag{6.28}$$

Formulons l'algorithme de résolution du problème. Soit x_0 une approximation initiale et supposons construits les points x_j, $j = 0, 1, \ldots, k$. Alors

$$x_{k+1} = x_k + \alpha_k p_k, \tag{6.29}$$

avec $p_k \equiv p(x_k)$. On choisit α_k égal à 2^{-i_0}, où i_0 est le premier des indices $i = 0, 1, \ldots,$ pour lequel

$$F(x_k + 2^{-i}p_k) \leqslant F(x_k) - 2^{-i}\varepsilon\|p_k\|^2, \quad \frac{1}{2} < \varepsilon < 1.$$

On a donc la condition

$$F(x_{k+1}) \leqslant F(x_k) - \alpha_k \varepsilon \| p_k \|^2. \tag{6.30}$$

Enonçons les conditions de convergence de l'algorithme.

LEMME 6.2. *$p(x) = 0$ si et seulement si on a au point x les conditions nécessaires de minimum de $F(x)$.*

On le démontre en rappelant les conditions nécessaires de minimum de $F(x)$ et en raisonnant comme pour le lemme 5.1.

THÉORÈME 6.4. *Etant donné des fonctions $f_i(x)$ continûment dérivables, un domaine*

$$\Omega = \{x\colon F(x) \leqslant F(x_0)\}$$

borné et $f_i'(x)$ vérifiant dans Ω la condition de Lipschitz de constante L, tout point limite x_* de la suite x_k, $k = 0, 1, \ldots$, vérifie les conditions nécessaires de minimum de $F(x)$ pour $x \in E^n$. Si $f_i(x)$ sont convexes, x_* est solution du problème.

DÉMONSTRATION. Comme pour le théorème 5.1, on trouve sans peine les estimations

$$f_i(x_k + \alpha p_k) \leqslant f_i(x_k) + \alpha (f_i'(x_k), p_k) + \alpha^2 L \| p_k \|^2,$$
$$i \in \mathcal{J}_\delta(x_k),$$
$$f_i(x_k + \alpha p_k) \leqslant F(x_k) - \delta + \alpha K \| p_k \|, \quad i \notin \mathcal{J}_\delta(x_k),$$

avec $K = \max\limits_{i,\, x \in \Omega} \| f_i'(x) \|$.

Utilisons la condition (voir (6.26))

$$(f_i'(x_k), p_k) \leqslant \beta_k - f_i(x_k), \qquad \beta_k = \beta(x_k),$$

ainsi que (6.28). La première estimation s'écrit alors

$$\begin{aligned} f_i(x_k + \alpha p_k) &\leqslant (1-\alpha) f_i(x_k) + \beta_k \alpha + \alpha^2 L \| p_k \|^2 \leqslant \\ &\leqslant F(x_k) - \alpha (F(x_k) - \beta_k) + \alpha^2 L \| p_k \|^2 \leqslant \\ &\leqslant F(x_k) - \frac{\alpha}{2} \| p_k \|^2 + \alpha^2 L \| p_k \|^2. \end{aligned}$$

Puisque

$$F(x_k) - \frac{\alpha}{2} \| p_k \|^2 \geqslant F(x_k) - \delta + \alpha K \| p_k \|$$

pour $0 \leqslant \alpha \leqslant \alpha_k^1$, $\alpha_k^1 = \dfrac{\delta}{\| p_k \| \left(K + \frac{1}{2} \| p_k \| \right)}$, on a

$$f_i(x_k + \alpha p_k) \leqslant F(x_k) - \frac{\alpha}{2} \| p_k \|^2 + \alpha^2 L \| p_k \|^2 \tag{6.31}$$

quand $0 \leqslant \alpha \leqslant \alpha_k^1$. Donc

$$F(x_k + \alpha p_k) \leqslant F(x_k) - \alpha \| p_k \|^2 \left[\frac{1}{2} - \alpha L\right], \quad 0 \leqslant \alpha \leqslant \alpha_k^1. \tag{6.32}$$

Si

$$0 \leqslant \alpha \leqslant \alpha_k, \ \bar{\alpha}_k = \min\left[1, \ \alpha_k^1, \ \frac{\varepsilon - \frac{1}{2}}{L}\right], \tag{6.33}$$

on a

$$F(x_k + \alpha p_k) \leqslant F(x_k) - \alpha \| p_k \|^2 \varepsilon.$$

Il en résulte de suite l'inégalité (6.30) lorsque

$$\alpha_k \geqslant \frac{1}{2} \bar{\alpha}_k \tag{6.34}$$

après un nombre fini de bipartitions de l'unité.

La condition (6.30) entraîne immédiatement $\alpha_k \| p_k \|^2 \to 0$. Or, cela signifie $\| p_k \| \to 0$. En effet, $\alpha_k^1 \geqslant \gamma > 0$ car $\| p(x) \|$ est majoré dans Ω en vertu de (6.27). Mais il découle de (6.33) et (6.34) que α_k est minoré par une constante positive.

Ainsi, $p_k \to 0$. Soit x_* un point limite de la suite. On dit, sans restreindre la généralité, que $x_k \to x_*$. Comme u_k^i, $i \in \mathcal{J}_\delta(x_k)$, sont positifs et donnent au total l'unité, on estime de plus, en posant $u_k^i = 0$, $i \notin \mathcal{J}_\delta(x_k)$, que $u_k^i \to \bar{u}^i$, $\bar{u}^i \geqslant 0$ et

$$\sum_{i=1}^{m} \bar{u}^i = 1. \tag{6.35}$$

Récrivons (6.27) et (6.26) pour les points x_k:

$$\begin{gathered} p_k + \sum_{i=1}^{m} u_k^i f_i'(x_k) = 0, \\ u_k^i ((f_i'(x_k), p_k) + f_i(x_k) - \beta_k) = 0, \quad i = 1, \ldots, m, \\ (f_i'(x_k), p_k) + f_i(x_k) \leqslant \beta_k, \quad i \in \mathcal{J}_\delta(x_k). \end{gathered} \tag{6.36}$$

La dernière inégalité entraîne

$$\beta_k \geqslant f_i(x_k) - K \| p_k \| = F(x_k) - K \| p_k \|$$

à condition de choisir $i \in \mathcal{J}_\delta(x_k)$ de façon que $f_i(x_k) = F(x_k)$. Mais (6.28) montre que $\beta_k \leqslant F(x_k) - \frac{1}{2} \| p_k \|^2$. Donc $\beta_k \to F(x_*)$. On obtient par passage à la limite dans (6.36):

$$\begin{gathered} \sum_{i=1}^{m} \bar{u}^i f_i'(x_*) = 0, \\ \bar{u}^i (f_i(x_*) - F(x_*)) = 0, \quad i = 1, \ldots, m, \\ \sum_{i=1}^{m} \bar{u}^i = 1, \ \bar{u}^i \geqslant 0. \end{gathered} \tag{6.37}$$

Ce qui constitue justement les conditions nécessaires pour que $F(x)$ atteigne son minimum en x_* (voir ch. I). Si $f_i(x)$ sont convexes, ces conditions sont également suffisantes, ce qui démontre le théorème. Quelle est l'estimation locale de la convergence de l'algorithme?

THÉORÈME 6.5. *Soient x_* un point de minimum de $F(x)$ et $f_i(x)$ des fonctions deux fois continûment dérivables. Soit de plus les gradients $f_i'(x_*)$, $i \in \mathcal{J}_0(x_*)$, où $\mathcal{J}_0(x_*) = \{i : f_i(x_*) = F(x_*)\}$, tels que les différences*

$$f_i'(x_*) - f_{i_0}'(x_*), \qquad i \neq i_0, \qquad i_0 \in \mathcal{J}_0(x_*),$$

soient linéairement indépendantes, les multiplicateurs $\bar{u}^i$ strictement supérieurs à 0 pour $i \in \mathcal{J}_0(x_*)$ et $(y, L''(x_*, \bar{u})\, y) > 0$ pour tous les $y \neq 0$. Ici

$$L(x, u) = \sum_{i=1}^{m} u^i f_i(x)$$

et $L''(x, u)$ est la matrice des dérivées secondes par rapport à x. Il existe alors pour un $\delta > 0$ suffisamment petit et $\alpha > 0$ un voisinage du point x_* tel que le processus

$$x_{k+1} = x_k + \alpha p(x_k), \qquad k = 0, 1, \ldots,$$

converge à partir de toute approximation initiale x_0 de ce voisinage et $\| x_* - x_k \| \leqslant Cq^k$, où $0 < q < 1$.

DÉMONSTRATION. Nous n'en donnerons que des grandes lignes vu que la démonstration est complètement analogue à celle du théorème 5.4 et s'y ramène en fait.

En posant

$$\tilde{\mathcal{J}}(x) = \{i \in \mathcal{J}_\delta(x) : \quad (f_i'(x), p(x)) + f_i(x) - \beta(x) = 0\},$$

on montre (voir lemme 5.2) que pour δ faible $\mathcal{J}_\delta(x) = \mathcal{J}_0(x_*) = \tilde{\mathcal{J}}(x)$ pour tous les x voisins de x_*. Il résulte donc de (6.26) et (6.27) que le vecteur $p(x)$ et les multiplicateurs de Lagrange u^i correspondants satisfont au système d'équations

$$\begin{gathered} p(x) + \sum_{i \in \mathcal{J}_0(x_*)} u^i f_i'(x) = 0, \\ (f_i'(x), p(x)) + f_i(x) = \beta(x), \quad i \in \mathcal{J}_0(x_*), \\ \sum_{i \in \mathcal{J}_0(x_*)} u^i = 1. \end{gathered} \tag{6.38}$$

Soit i_0 un indice de $\mathcal{J}_0(x_*)$ et

$$\tilde{f}_i(x) = f_i(x) - f_{i_0}(x), \qquad \tilde{f}_0(x) = f_{i_0}(x).$$

Le système (6.38) équivaut alors à

$$p(x)+\tilde{f}_0'(x)+\sum_{i\in\tilde{\mathcal{J}}_0(x_*)} u^i\tilde{f}_i'=0,$$

$$(\tilde{f}_i'(x),\ p(x))+\tilde{f}_i(x)=0,\quad i\in\tilde{\mathcal{J}}_0(x_*), \tag{6.39}$$

$$\mathcal{J}_0(x_*)=\mathcal{J}_0(x_*)\setminus\{i_0\}.$$

Or, on constate une équivalence totale entre celui-ci et les systèmes (5.51.1) et (5.51.2). Comme la démonstration du théorème 5.4 s'est ramenée à l'étude des propriétés de $p(x)$, solution de (5.51.1), (5.51.2), il en résulte que la démonstration ultérieure du théorème 6.5 se réduit à une vérification des conditions du théorème 5.4. Mais on vérifie aisément que les hypothèses de celui-là les garantissent complètement pour les fonctions $\tilde{f}_i$, et le théorème se trouve démontré.

L'analogue du théorème 5.5 est le

THÉORÈME 6.6. *Si, dans les conditions du théorème* 6.5, *les indices de l'ensemble* $\mathcal{J}_0(x_*)$ *sont au nombre de* $n+1$, *pour* δ *petit la convergence du processus*

$$x_{k+1}=x_k+p(x_k) \tag{6.40}$$

vers le point x_* *est quadratique.*

DÉMONSTRATION. Dans le cas considéré le vecteur $p(x)$ est défini de façon unique par le système

$$(\tilde{f}_i'(x),\ p(x))+\tilde{f}_i(x)=0,\quad i\in\tilde{\mathcal{J}}_0(x_*),$$

car il y a indépendance linéaire des $\tilde{f}_i'(x)$, $i\in\tilde{\mathcal{J}}_0(x_*)$, pour x voisins par hypothèse de x_*. Le processus (6.40) n'est donc rien autre que la méthode de Newton pour le système

$$\tilde{f}_i(x)=0,\qquad i\in\tilde{\mathcal{J}}_0(x_*) \tag{6.41}$$

qui a la convergence quadratique au voisinage de x_* en vertu du théorème 6.1 et de la remarque 1 du n° 3 de ce paragraphe. Notons que le point x_* vérifie (6.41) car $f_i(x_*)=F(x_*)$, $i\in\mathcal{J}_0(x_*)$, et donc

$$\tilde{f}_i(x_*)=f_i(x_*)-f_{i_0}(x_*)=0,\qquad i\in\tilde{\mathcal{J}}_0(x_*).$$

§ 7. Accélération locale de la convergence

On a établi au § 5 que la méthode de linéarisation converge en général en progression géométrique. Cette vitesse peut ne pas suffire dans certains problèmes, d'où la nécessité de l'accélérer.

Ce paragraphe fait connaître des techniques qui permetten de le faire si l'on dispose d'une approximation suffisamment bonne de

la solution. Cette dernière restriction désavantage le processus mais on ne connaît malheureusement pas de méthodes avec lesquelles on construirait, à partir d'une approximation de départ quelconque, un processus ayant une convergence asymptotiquement superlinéaire comme c'est le cas du problème de minimisation de fonctions sans contraintes.

Les méthodes ci-dessous s'inspirent de l'idée suivante. Le problème de minimisation se ramène à un système d'équations non linéaires qu'on résout par la méthode de Newton ou par l'une de ses versions. Le procédé exposé à la fin du paragraphe part directement de cette idée, i.e. on écrira des conditions nécessaires de minimum et on appliquera aux équations obtenues la méthode de Newton. Ce procédé présente plusieurs défauts dont le principal est la nécessité de calculer les dérivées secondes des fonctions originelles. Il n'est donc utilisé que dans des problèmes où ce calcul se fait aisément.

L'idée de base d'une autre méthode est que le point x_* est solution du problème de minimisation (5.1) si et seulement si il vérifie l'équation $p(x_*) = 0$ avec $p(x)$ solution du problème auxiliaire (5.4). On exposera une technique qui permet de résoudre le système d'équations non linéaires sans calculer les dérivées et qui converge, nous l'avons dit, à partir d'une approximation initiale suffisamment bonne.

1. Position du problème. Formules fondamentales. Soit à résoudre le système d'équations

$$p(x) = 0, \tag{7.1}$$

où $p(x)$ est un vecteur de composantes $p^i(x)$, $i = 1, \ldots, n$, $x \in E^n$. Notons que $p(x)$ est une fonction vectorielle arbitraire qui n'est pas liée pour le moment à un problème de programmation mathématique.

Soit x_* solution du problème (7.1). Supposons que $p(x)$ est dérivable dans le voisinage de x_* et la matrice des dérivées

$$p'(x) = \left\{\frac{\partial p^i(x)}{\partial x^j}\right\}_{\substack{i=1,\ldots,n\\ j=1,\ldots,n}}$$

vérifie la condition de Lipschitz, i.e.

$$\| p'(x) - p'(y) \| \leqslant L \| x - y \|,$$

toutes les normes étant euclidiennes.

On estime sans nuire à la généralité que $x_* = 0$. Notons

$$p'(0) = A, \quad \omega(x) = p(x) - Ax,$$
$$\omega(x, y) = \frac{1}{\|x-y\|}[p(x) - p(y) - A(x-y)]. \tag{7.2}$$

Supposons la matrice A régulière de sorte qu'on a les estimations

$$m \| x \| \leqslant \| Ax \| \leqslant M \| x \|, \tag{7.3}$$

où $M \geqslant m > 0$.

LEMME 7.1. *On a les estimations*

$$\| \omega (x) \| \leqslant C_1 \| x \|^2, \quad \| \omega (x, y) \| \leqslant C_2 \max \{\| x \|, \| y \|\}.$$

DÉMONSTRATION. Soit $p^{i\prime} (x)$ gradient de $p^i (x)$. On a alors selon la formule de Taylor

$$p^i (x) = p^i (0) + (p^{i\prime} (0), x) + (p^{i\prime} (z) - p^{i\prime} (0), x),$$

avec $z = \theta x$, $0 \leqslant \theta \leqslant 1$. Utilisons le fait que $p^i (0) = 0$ et la condition de Lipschitz pour $p' (x)$, il vient de suite

$$\| p^i (x) - (p^{i\prime} (0), x) \| \leqslant L \| x \|^2, \quad i = 1, \ldots, n.$$

D'où la formule

$$\| \omega (x) \| = \| p (x) - p' (0) x \| \leqslant C_1 \| x \|^2.$$

Ensuite

$$p^i (y) = p^i (x) + (p^{i\prime} (x), y - x) + (p^{i\prime} (z) - p^{i\prime} (x), y - x),$$

où $z = \theta x + (1 - \theta) y$, $0 \leqslant \theta \leqslant 1$. Donc

$$p^i (y) - p^i (x) - (p^{i\prime} (0), y - x) = (p_i' (x) - p_i' (0), y - x) + \\ + (p^{i\prime} (z) - p^{i\prime} (x), y - x),$$

d'où l'on obtient par de simples transformations et compte tenu de la condition de Lipschitz :

$$\begin{aligned} | p^i (y) - p^i (x) - (p^{i\prime} (0), y - x) | &\leqslant L \| x \| \| y - x \| + L \| z - x \| \| y - x \| = \\ &= L \| y - x \| (\| x \| + (1 - \theta) \| y - x \|) \leqslant \\ &\leqslant L \| y - x \| ((2 - \theta) \| x \| + (1 - \theta) \| y \|) \leqslant \\ &\leqslant 3L \| y - x \| \max \{\| x \|, \| y \|\}. \end{aligned}$$

La dernière inégalité entraîne de suite la deuxième affirmation du lemme.

Supposons construits les points $x_1, x_2, \ldots, x_n$ et soit $p (x_k) \neq 0$, $k = 1, \ldots, n$, et e_k les vecteurs unités du k-ième axe.

Posons

$$\begin{aligned} y_k &= x_k + \| p (x_k) \| e_k, \\ r_k &= y_k - x_k = \| p (x_k) \| e_k, \\ z_k &= p (y_k) - p (x_k), \quad k = 1, \ldots, n. \end{aligned}$$

Introduisons une quantité qui mesure l'indépendance linéaire d'une famille quelconque de vecteurs b_k, $k = 1, \ldots, n$. Posons

$$\Delta(b_1, \ldots, b_n) = \min_{\sum_{i=1}^{n} |\alpha_i| = 1} \left\| \sum_{i=1}^{n} \alpha_i \frac{b_i}{\| b_i \|} \right\|.$$

On voit sans peine que $\Delta(b_1, \ldots, b_n) > 0$ si et seulement si les vecteurs $b_1, \ldots, b_n$ sont linéairement indépendants. Notons de plus que

$$\Delta(e_1, \ldots, e_n) = \frac{1}{\sqrt{n}}.$$

LEMME 7.2. *Il existe un voisinage du point $x_* = 0$ tel que*

$$\Delta(z_1, \ldots, z_n) \geqslant \gamma > 0$$

si et seulement si $x_1, \ldots, x_n$ se situent dans ce voisinage.

DÉMONSTRATION. On a par définition de $\omega(x, y)$

$$z_k = Ar_k + \omega(y_k, x_k) \| r_k \| = \| p(x_k) \| (Ae_k + \omega(y_k, x_k)).$$

Donc

$$\frac{z_k}{\| z_k \|} = \frac{Ae_k + \omega(y_k, x_k)}{\| Ae_k + \omega(y_k, x_k) \|}.$$

Si $x_k \to 0$, on a

$$\frac{z_k}{\| z_k \|} \to \frac{Ae_k}{\| Ae_k \|}.$$

On constate cependant sans peine que $\Delta(z_1, \ldots, z_n)$ dépend continûment de $z_k \| z_k \|^{-1}$. On a donc, pour des x_k suffisamment proches de 0,

$$\Delta(z_1, \ldots, z_n) \geqslant \frac{1}{2} \Delta(Ae_1, \ldots, Ae_n).$$

Mais $\Delta(Ae_1, \ldots, Ae_n) > 0$ car les vecteurs Ae_k, $k = 1, \ldots, n$, sont tout simplement les colonnes de la matrice A, et, étant donné la régularité de celle-ci, ses colonnes sont linéairement indépendantes. Ainsi,

$$\Delta(z_1, \ldots, z_n) \geqslant \frac{1}{2} \Delta(Ae_1, \ldots, Ae_n) > 0$$

pour tous les x_k d'un voisinage de $x_* = 0$, c.q.f.d.

Soit $\delta > 0$ le rayon d'un voisinage de zéro où sont vrais les lemmes 7.1 et 7.2, et soient $x_1, \ldots, x_n$ des points choisis dans ce voisinage. Trouvons les quantités β_i, $i = 1, \ldots, n$, à partir du

système d'équations

$$-p(x_n)=\sum_{k=1}^{n}\beta_k z_k. \tag{7.4}$$

Par suite du lemme 7.2, il possède une solution. Posons

$$x_{n+1}=x_n+\sum_{k=1}^{n}\beta_k r_k \tag{7.5}$$

et évaluons la norme de x_{n+1}. Comme

$$p(x_n)=Ax_n+\omega(x_n), \tag{7.6}$$
$$z_k=Ar_k+\omega(y_k,\ x_k)\,\|r_k\|,$$

(7.4) fournit

$$-Ax_n-\omega(x_n)=\sum_{k=1}^{n}\beta_k Ar_k+\sum_{k=1}^{n}\beta_k\omega(y_k,\ x_k)\,\|r_k\|$$

ou

$$Ax_{n+1}=-\omega(x_n)-\sum_{k=1}^{n}\beta_k\omega(y_k,\ x_k)\,\|r_k\|.$$

Il découle de la dernière égalité

$$m\,\|x_{n+1}\|\leqslant\|Ax_{n+1}\|\leqslant\|\omega(x_n)\|+\sum_{k=1}^{n}|\beta_k|\,\|r_k\|\,\|\omega(y_k,\ x_k)\|. \tag{7.7}$$

Mais en vertu du lemme 7.1

$$\|\omega(y_k,\ x_k)\|\leqslant C_2\max\{\|y_k\|,\ \|x_k\|\}=C_2\max\{\|(x_k+$$
$$+\|p(x_k)\|\,e_k)\|,\ \|x_k\|\}\leqslant C_2(\|x_k\|+\|p(x_k)\|).$$

Dans le voisinage considéré

$$\|p(x_k)\|=\|Ax_k+\omega(x_k)\|\leqslant M\,\|x_k\|+C_1\,\|x_k\|^2.$$

Aussi

$$\|\omega(y_k,\ x_k)\|\leqslant C_2(1+M+C_1\delta)\,\|x_k\|\equiv C_3\,\|x_k\|.$$

L'estimation (7.7) se récrit moyennant cette inégalité :

$$\|x_{n+1}\|\leqslant\frac{1}{m}\left[C_1\,\|x_n\|^2+C_3\max_{1\leqslant k\leqslant n}\|x_k\|\sum_{k=1}^{n}|\beta_k|\,\|r_k\|\right]. \tag{7.8}$$

Notons que

$$\left\|\sum_{k=1}^{n}\beta_k z_k\right\|=\left\|\sum_{k=1}^{n}\frac{\beta_k\,\|z_k\|}{\sum\limits_{i=1}^{n}|\beta_i|\,\|z_i\|}\,\frac{z_k}{\|z_k\|}\right\|\left(\sum_{i=1}^{n}|\beta_i|\,\|z_i\|\right)\geqslant$$

$$\geqslant\Delta(z_1,\ \ldots,\ z_n)\left(\sum_{k=1}^{n}|\beta_k|\,\|z_k\|\right)$$

Compte tenu de (7.4), on a

$$\| p(x_n) \| \geqslant \Delta(z_1, \ldots, z_n) \Big(\sum_{k=1}^{n} |\beta_k| \, \| z_k \| \Big). \tag{7.9}$$

Ensuite,

$$\| z_k \| = \| (Ar_k + \omega(y_k, x_k) \| r_k \|) \| \geqslant \| Ar_k \| - \| r_k \| \, \| \omega(y_k, x_k) \| \geqslant$$
$$\geqslant \| r_k \| (m - C_3 \| x_k \|) \geqslant \| r_k \| (m - C_3 \max_{1 \leqslant k \leqslant n} \| x_k \|).$$

Donc

$$\sum_{k=1}^{n} |\beta_k| \, \| z_k \| \geqslant \Big(\sum_{k=1}^{n} |\beta_k| \, \| r_k \| \Big) (m - C_3 \max_{1 \leqslant k \leqslant n} \| x_k \|).$$

Si

$$\max_{1 \leqslant k \leqslant n} \| x_k \| < m/C_3, \tag{7.10}$$

l'inégalité (7.9) fournit alors

$$\sum_{k=1}^{n} |\beta_k| \, \| r_k \| \leqslant \frac{\| p(x_n) \|}{\Delta(z_1, \ldots, z_n)(m - C_3 \max_{1 \leqslant k \leqslant n} \| x_k \|)}. \tag{7.11}$$

Vu que

$$\| p(x_n) \| = \| Ax_n + \omega(x_n) \| \leqslant M \| x_n \| + C_1 \| x_n \|^2 \leqslant C_4 \| x_n \|,$$

(7.8) se récrit finalement compte tenu de (7.11) et du lemme 7.2:

$$\| x_{n+1} \| \leqslant \| x_n \| \frac{1}{m} \left[C_1 \| x_n \| + \frac{C_3 \max_{1 \leqslant k \leqslant n} \| x_k \| \, C_4}{\gamma (m - C_3 \max_{1 \leqslant k \leqslant n} \| x_k \|)} \right]. \tag{7.12}$$

Enonçons le résultat obtenu sous forme du

LEMME 7.3. *Si les points* x_k, $k = 1, \ldots, n$, *sont choisis dans un voisinage de* $x_* = 0$ *qui vérifie les conditions des lemmes* 7.1, 7.2 *et l'inégalité* (7.10), *on a l'estimation* (7.12).

2. Algorithme. Formulons l'algorithme de résolution du système (7.1).

Choisissons les points de départ $x_1, x_2, \ldots, x_n$. Les points $x_1, \ldots, x_n, \ldots, x_k$ étant supposés construits, x_{k+1} est généré par la formule suivante:

$$x_{k+1} = x_k + \sum_{i=1}^{n} \beta_i^k r_{k-n+i}, \tag{7.13}$$

où

$$r_j = \| p(x_j) \| \, e_{m(j)}, \; y_j = x_j + r_j, \; z_j = p(y_j) - p(x_j),$$

et les quantités β_i^k, $i = 1, \ldots, n$, se définissent par le système

$$-p(x_k) = \sum_{i=1}^{n} \beta_i^k z_{k-n+i}. \tag{7.14}$$

Le calcul de l'indice $m(j)$ obéit à la règle suivante : si $j = ln + p$, $1 \leqslant p \leqslant n - 1$, avec l entier, alors $m(j) = p$, et si $j = ln$, on a $m(j) = n$.

Ainsi, les vecteurs $r_1, r_2, \ldots, r_k$ sont proportionnels aux vecteurs unités des axes de coordonnées pris dans l'ordre cyclique.

Comme il résulte des formules citées, le schéma de l'algorithme est assez simple. Il comprend à chaque pas le calcul de $p(x)$ aux points x_k et y_k et la résolution du système (7.14).

THÉORÈME 7.1. *Soit $\delta_0 > 0$ tel que les x, $\| x \| \leqslant \delta_0$, vérifient les conditions des lemmes 7.1, 7.2 et les inégalités*

$$m - C_3 \| x \| \geqslant \frac{m}{2}, \tag{7.15}$$

$$\frac{\delta_0}{m}\left[C_1 + \frac{2C_3C_4}{m\gamma}\right] < 1. \tag{7.16}$$

Etant donné $x_1, \ldots, x_n$ choisis de façon que $\| x_k \| \leqslant \delta_0$, $k = 1, \ldots, n$, l'algorithme décrit converge vers la solution x_* des équations (7.1) et la convergence est superlinéaire.

DÉMONSTRATION. Montrons en premier lieu que $\| x_k \| \leqslant \delta_0$ pour tous les points x_k générés par l'algorithme. En effet, si $x_1, \ldots, x_k$ se trouvent dans le δ_0-voisinage de x_*, les points $x_{k-n+1}, x_{k-n+2}, \ldots, x_k$ remplissent les conditions du lemme 7.3 et on a donc l'inégalité suivante qui est analogue à (7.12) :

$$\| x_{k+1} \| \leqslant \| x_k \| \frac{1}{m}\left[C_1 \| x_k \| + \frac{C_3C_4 \max\limits_{1 \leqslant i \leqslant n} \| x_{k-n+i} \|}{\gamma(m - C_3 \max\limits_{1 \leqslant i \leqslant n} \| x_{k-n+i} \|)}\right],$$

d'où, compte tenu de (7.15),

$$\| x_{k+1} \| \leqslant \| x_k \| \max_{1 \leqslant i \leqslant n} \| x_{k-n+i} \| C_5, \tag{7.17}$$

avec

$$C_5 = \frac{1}{m}\left[C_1 + \frac{2C_3C_4}{\gamma m}\right].$$

Mais $\| x_{k-n+i} \| \leqslant \delta_0$ par hypothèse. Aussi

$$\| x_{k+1} \| \leqslant \| x_k \| \delta_0 C_5 \leqslant \| x_k \| \leqslant \delta_0,$$

c.q.f.d. A condition de noter $q_0 = \delta_0 C_5$, la dernière inégalité implique de plus

$$\| x_{k+1} \| \leqslant q_0 \| x_k \|. \tag{7.18}$$

Comme $q_0 < 1$ en vertu de (7.16), il en découle l'estimation $\| x_k \| \leqslant q_0^{k-n} \| x_n \|$, i.e. $x_k \to 0$, ce qui démontre la première affirmation du théorème.

On obtient ensuite à partir de (7.17)

$$\frac{\| x_{k+1} \|}{\| x_k \|} \leqslant C_5 \max_{1 \leqslant i \leqslant n} \| x_{k-n+i} \|. \tag{7.19}$$

Etant donné que x_k tend vers zéro, la dernière estimation signifie que

$$\frac{\| x_{k+1} \|}{\| x_k \|} \to 0.$$

Cette relation montre que $x_k \to 0$ plus vite qu'en progression géométrique. Le théorème se trouve démontré.

Evaluons la convergence d'une façon plus précise. Posons $v_k = C_5 \| x_k \|$. (7.17) se récrit alors

$$v_{k+1} \leqslant v_k \max_{1 \leqslant i \leqslant n} v_{k-n+i}. \tag{7.20}$$

Posons maintenant $\tilde{v}_j = \max_{1 \leqslant i \leqslant n} v_i$, $j = 1, \ldots, n$, et définissons $\tilde{v}_k$, $k > n$, par la formule de récurrence

$$\tilde{v}_{k+1} = \tilde{v}_k \max_{1 \leqslant i \leqslant n} \tilde{v}_{k-n+i}. \tag{7.21}$$

On voit aisément que $v_k \leqslant \tilde{v}_k$ pour tous les k. Puisque

$$v_i = C_5 \| x_i \| \leqslant C_5 \delta_0 = q_0 < 1, \quad i = 1, \ldots, n,$$

on a $\tilde{v}_i \leqslant q_0 < 1$, $i = 1, \ldots, n$, d'où la décroissance monotone de la suite $\tilde{v}_k$. On le démontre fort simplement par récurrence sur k. Il en résulte $\max_{1 \leqslant i \leqslant n} \tilde{v}_{k-n+i} = \tilde{v}_{k-n+1}$, et (7.21) se récrit comme suit :

$$\tilde{v}_{k+1} = \tilde{v}_k \tilde{v}_{k-n+1}. \tag{7.22}$$

Posons $w_k = \ln \tilde{v}_k$. On a alors

$$\begin{aligned} w_{k+1} &= w_k + w_{k-n+1}, \quad k \geqslant n, \\ w_k &= \ln \tilde{v}_k, \quad k = 1, \ldots, n. \end{aligned} \tag{7.23}$$

Les résultats d'Ostrowski ([91], théorèmes 12.1 et 12.2) entraînent

$$\frac{w_{k+1}}{w_k} \to \lambda_0, \tag{7.24}$$

λ_0 étant la plus grande racine positive de l'équation

$$\varphi(\lambda) = \lambda^n - \lambda^{n-1} - 1 = 0. \tag{7.25}$$

Comme $\varphi(1) = -1 < 0$ et $\varphi(\lambda) > 0$ pour λ importants, on a $\lambda_0 > 1$.

Il découle de (7.24) que, quel que soit $\varepsilon > 0$, $\lambda_0 - \varepsilon > 1$, on trouve un numéro $k(\varepsilon)$ tel que $\frac{w_{k+1}}{w_k} \geqslant \lambda_0 - \varepsilon$ ou $\frac{\ln \tilde{v}_{k+1}}{\ln \tilde{v}_k} \geqslant \lambda_0 - \varepsilon$.

Puisque $\ln \tilde{v}_k < 0$ $(\tilde{v}_k < q_0 < 1)$, on a pour $k \geqslant k(\varepsilon)$

$$\ln \tilde{v}_{k+1} \leqslant (\lambda_0 - \varepsilon) \ln \tilde{v}_k = \ln \tilde{v}_k^{(\lambda_0 - \varepsilon)}$$

ou $\tilde{v}_{k+1} \leqslant \tilde{v}_k^{(\lambda_0-\varepsilon)}$, $k \geqslant k(\varepsilon)$. La dernière formule entraîne $\tilde{v}_k \leqslant \leqslant \tilde{v}_{k(\varepsilon)}^{(\lambda_0-\varepsilon)^{(k-k(\varepsilon))}}$. Mais la suite $\tilde{v}_k$ est monotone décroissante et $\tilde{v}_n \leqslant q_0 < 1$. Donc

$$\tilde{v}_k \leqslant q_0^{(\lambda_0-\varepsilon)^{(k-k(\varepsilon))}}, \quad k \geqslant k(\varepsilon). \tag{7.26}$$

THÉORÈME 7.2. *Dans les conditions du théorème 7.1 il existe, pour tout* $\varepsilon > 0$, $\lambda_0 - \varepsilon > 1$, *où* λ_0 *est la plus grande racine de* $\lambda^n - \lambda^{n-1} - 1 = 0$, *un numéro* $k(\varepsilon)$ *tel que n'importe quel* $k \geqslant k(\varepsilon)$ *vérifie l'inégalité*

$$\|x_k\| \leqslant \frac{1}{C_5} q_0^{(\lambda_0-\varepsilon)^{(k-k(\varepsilon))}}, \quad q_0 < 1. \tag{7.27}$$

DÉMONSTRATION. Rappelons que $v_k = C_5 \|x_k\|$, $v_k \leqslant \tilde{v}_k$. Le résultat cherché résulte de suite de ces inégalités et de (7.26).

3. Aspects numériques. Application à un problème de programmation mathématique. L'algorithme du numéro précédent est assez simple. Il n'exige à chaque pas que le calcul du vecteur $p(x)$ aux points x_k et y_k et la résolution du système d'équations linéaires (7.14). En désignant par Z_k la matrice de colonnes z_{k-n+i}, $i = 1, \ldots, n$, on peut écrire (7.14) sous la forme $Z_k \beta^k = -p(x_k)$, avec β^k un vecteur-colonne de composantes β_i^k, $i = 1, \ldots, n$.

Selon l'algorithme, les matrices Z_k et Z_{k-1} ne diffèrent que par une colonne : la colonne z_{k+1} est substituée à z_k et $z_{k-n+i+1}$, $i \leqslant n-1$, à z_{k-n+i}, $i \leqslant n-1$. On calcule donc β_i^{k+1} par les procédés décrits aux n°s 4, 5, § 3, ch. II. Il n'est pas inutile de dire que ces procédés s'accompagnent de la propagation de l'erreur de calcul. Si la recherche de $p(y_k)$ demande donc un effort mathématique sensiblement plus important que la résolution du système (7.14), on recourt à un programme standard de résolution de ce dernier pour trouver β^k et on se passe de formules de récurrence.

Revenons au problème (5.1)-(5.2) du § 5. Aux termes du lemme 5.1, il suffit, pour trouver le minimum local, de résoudre l'équation $p(x) = 0$, où $p(x)$ est solution de (5.4). Sous les hypothèses (suffisamment naturelles) du théorème 5.4, les conditions du théorème 7.1 sont vérifiées, en vertu des lemmes 5.2, 5.4, dans un voisinage suffisamment petit de la solution x_*. Avec l'algorithme exposé dans ce paragraphe on peut donc accélérer la convergence de la méthode de linéarisation. On prend alors pour $p(x)$ le vecteur $p(x)$ qui est solution du problème (5.4).

4. Un problème de minimisation avec contraintes égalité. Considérons le problème de la recherche du minimum de $f_0(x)$ sous les contraintes

$$f_i(x) = 0, \quad i = 1, \ldots, m. \tag{7.28}$$

Soit x_* solution du problème et supposons vraies les hypothèses suivantes.

a) Les fonctions $f_i(x)$ sont deux fois continûment dérivables et leurs dérivées secondes vérifient la condition de Lipschitz.

b) Au point x_* les gradients $f_i'(x_*)$, $t = 1, \ldots, m$, sont linéairement indépendants si bien que les conditions nécessaires de minimum en x_* sont vérifiées sous forme régulière (voir ch. I, § 4). Il existe donc des multiplicateurs de Lagrange u^i, $i = 1, \ldots, m$, tels que

$$\begin{gathered} f_0'(x_*) + \sum_{i=1}^{m} u^i f_i'(x_*) = 0, \\ f_i(x_*) = 0, \quad i = 1, \ldots, m. \end{gathered} \tag{7.29}$$

c) On a les conditions suffisantes de minimum local, i.e. $(y, L''(x_*, u)\, y) > 0$ si $y \neq 0$ et $(f_i'(x_*), y) = 0$, $i = 1, \ldots, m$. Ici $L(x, u) = f_0(x) + \sum_{i=1}^{m} u^i f_i(x)$ et $L''(x, u)$ est la matrice des dérivées secondes de $L(x, u)$ par rapport à x.

THEOREME 7.3. *Etant donné les conditions* a)-c), *les suites* x_k, u_k^i, $i = 1, \ldots, m$, $k = 0, 1, \ldots$, *calculées moyennant les formules de récurrence*

$$L''(x_k, u_k)\, p_k + \sum_{i=1}^{m} \Delta u_k^i f_i'(x_k) + L'(x_k, u_k) = 0, \tag{7.30}$$

$$(f_i'(x_k), p_k) + f_i(x_k) = 0, \quad i = 1, \ldots, m,$$

$$\begin{aligned} x_{k+1} &= x_k + p_k, \\ u_{k+1}^i &= u_k^i + \Delta u_k^i, \quad i = 1, \ldots, m, \end{aligned} \tag{7.31}$$

convergent vers x_* *et* u^i *respectivement à partir de toute approximation initiale* x_0, u_0^i, $i = 1, \ldots, m$, *suffisamment voisine de la solution* x_*, u^i, $i = 1, \ldots, m$, *et la convergence est quadratique.*

DEMONSTRATION. Le processus défini par les formules (7.30), (7.31) n'est rien d'autre que le processus généré par la méthode de Newton appliquée au système (7.29). Il suffit donc de vérifier, comme l'indique la remarque 1 du n° 3, § 6, que la matrice des dérivées premières des premiers membres de (7.29) par rapport à tous les arguments x et u^i est régulière par rapport à la solution.

Désignons par $f'(x)$ la matrice de lignes $f_i'(x)$, $i = 1, \ldots, m$; on voit sans peine que la matrice des dérivées premières des pre-

miers membres de (7.29) s'écrit par blocs

$$\underbrace{\begin{pmatrix} L''(x_*, u) & f'^*(x_*) \\ f'(x_*) & 0 \end{pmatrix}}_{n+m} \Big\} n+m.$$

Pour s'assurer de sa régularité, il suffit de montrer que le système homogène

$$\begin{aligned} L''(x_*, u)\,\bar{y} + f'^*(x_*)\,\bar{u} &= 0, \\ f'(x_*)\,\bar{y} &= 0 \end{aligned} \tag{7.32}$$

n'admet d'autre solution que 0. Ici $\bar{y} \in E^n$ et $\bar{u}$ un vecteur de composantes $\bar{u}^i$, $i = 1, \ldots, m$. Ainsi, soit $\bar{y}$, $\bar{u}$ solution du système (7.32). Multiplions scalairement par $\bar{y}$ la première équation (7.32), il vient en vertu de la seconde

$$(\bar{y}, L''(x_*, u)\,\bar{y}) + (\bar{y}, f'^*(x_*)\,\bar{u}) =$$
$$= (\bar{y}, L''(x_*, u)\,\bar{y}) + (f'(x_*)\,\bar{y}, \bar{u}) = (\bar{y}, L''(x_*, u)\,\bar{y}) = 0.$$

Mais, par suite de l'hypothèse c), la dernière relation montre que $\bar{y} = 0$. La première relation (7.32) prend donc la forme

$$f'^*(x_*)\,\bar{u} \equiv \sum_{i=1}^{m} \bar{u}^i f'_i(x_*) = 0,$$

ce qui n'est possible que si $\bar{u}^i = 0$, $i = 1, \ldots, m$, du moment que les vecteurs $f'_i(x_*)$ sont linéairement indépendants par hypothèse b).

Ainsi, on a montré qu'ont lieu les conditions de convergence de la méthode de Newton, ce qui démontre du même coup le théorème.

§ 8. Méthode de pénalisation

Il s'agit d'une méthode de résolution d'un problème de programmation mathématique parmi les plus simples et le mieux connues. Son idée de base consiste à ramener approximativement le problème de minimisation avec contraintes à la recherche du minimum d'une fonction sans contraintes. La fonction auxiliaire est alors choisie de façon qu'elle coïncide avec la fonction à minimiser à l'intérieur du domaine admissible et qu'elle augmente rapidement à son extérieur.

Admettons qu'on a à minimiser $f_0(x)$, $x \in E^n$, sous les contraintes

$$f_i(x) \leqslant 0, \quad i = 1, \ldots, m. \tag{8.1}$$

Toutes les fonctions $f_i(x)$, $i = 0, 1, \ldots, m$, sont continues.

Soit

$$\varphi_0(t)=\begin{cases} t^2, & t\geqslant 0,\\ 0, & t<0;\end{cases}\qquad \varphi_1(t)=\begin{cases} t, & t\geqslant 0,\\ 0, & t<0.\end{cases} \tag{8.2}$$

Formons la fonction

$$\psi(x, r)=r\sum_{i=1}^{m}\varphi_0(f_i(x)). \tag{8.3}$$

On voit aisément que

$$\psi(x, r)=0,\quad x\in\Omega,$$

où

$$\Omega=\{x: f_i(x)\leqslant 0,\ i=1,\ \ldots,\ m\}.$$

Dans le cas de x n'appartenant pas à Ω, on a $\psi(x, r)>0$ et $\psi(x, r)\to$ $\to+\infty$ pour $r\to+\infty$. Le problème auxiliaire est maintenant de minimiser la fonction

$$F(x, r)\equiv f_0(x)+\psi(x, r). \tag{8.4}$$

On s'attend naturellement que la solution $x(r)$ de ce problème sera proche de celle du problème initial. Plus bas on en formulera les conditions exactes.

Notons qu'on ne choisit nécessairement pas $\psi(x, r)$ comme indiqué plus haut. Il suffit que cette fonction jouisse de certaines propriétés générales garantissant la convergence de la méthode. Les propriétés de celle-ci diffèrent selon le choix des $\psi(x, r)$. En particulier, si l'on pose

$$\psi(x, r)=r\max_{1\leqslant i\leqslant r}\varphi_1(f_i(x)),$$

la méthode de linéarisation exposée au § 5 peut s'interpréter comme méthode de la recherche du minimum de (8.4). Il n'est pas besoin dans ce cas, selon le § 5, de faire tendre r vers l'infini. La fonction $F(x, r)$ sera cependant non régulière.

Dans le cas général, on construit $F(x, r)$ de façon qu'elle soit régulière et qu'on puisse appliquer l'une des méthodes à convergence rapide du chapitre II. Malheureusement, il faut alors faire tendre r vers l'infini, d'où certaines difficultés cachées qui, selon les auteurs, déprécient beaucoup la méthode de pénalisation. Nous discuterons plus loin ces difficultés et nous exposerons brièvement une autre technique dont l'idée s'approche de celle de la méthode étudiée.

1. Justification de la méthode de pénalisation. Supposons qu'une fonction continue $\psi(x, r)$ possède les propriétés suivantes:

1) $\psi(x, r)=0$ si $x\in\Omega$; $\psi(x, r)>0$ si $x\notin\Omega$; $\psi(x_k, r_k)\to+\infty$ si $x_k\to x_0$, $x_0\notin\Omega$, $r_k\to+\infty$;

2) $\psi(x, r)$ croît de façon monotone avec r.

THÉORÈME 8.1. *Etant donné un ensemble*

$$\Omega_C(r) = \{x\colon F(x, r) \leqslant C\},$$
$$F(x, r) = f_0(x) + \psi(x, r),$$

compact, la fonction $F(x, r)$ atteint non minimum $m(r)$ par rapport à tous les x en un point $x(r)$, $m(r) \leqslant m$, avec

$$m = \min_{x \in \Omega} f_0(x), \quad m(r) \to m,$$

et $m(r)$ croît monotonement avec r. De plus, si $x(r_k) \to x_0$, $k \to \infty$, $r_k \to \infty$, alors x_0 est solution du problème de départ (8.1).

DÉMONSTRATION. Soit $\bar{x}$ un point de Ω, $\bar{C} = f_0(\bar{x})$. L'ensemble $\bar{\Omega}$ des $x \in \Omega$ tels que $f_0(x) \leqslant \bar{C}$ est alors une partie fermée de l'ensemble compact $\Omega_{\bar{C}}(r)$. En effet, pour $x \in \bar{\Omega}$ on a, en vertu des propriétés de $\psi(x, r)$,

$$f_0(x) + \psi(x, r) = f_0(x) \leqslant \bar{C},$$

i.e. $x \in \Omega_{\bar{C}}(r)$. Il est cependant clair que le minimum de $f_0(x)$ sur Ω doit appartenir au sous-ensemble $\bar{\Omega}$. On recherche donc le minimum de la fonction continue $f_0(x)$ sur l'ensemble compact $\bar{\Omega}$. Comme une fonction continue atteint son minimum sur un ensemble compact, il en résulte la résolubilité du problème (8.1). En raisonnant de façon analogue on montre que la fonction $F(x, r)$ atteint son minimum $m(r)$ en un point $x(r)$.

Soit x_* un point de minimum de $f_0(x)$ dans Ω. Alors

$$F(x_*, r) = f_0(x_*) + \psi(x_*, r) = f_0(x_*),$$

car $x_* \in \Omega$ et $\psi(x, r) = 0$ pour $x \in \Omega$. Donc

$$\min_x F(x, r) = m(r) \leqslant m,$$

i.e. $x(r) \in \Omega_m(r)$.

Considérons les ensembles

$$\Omega_m(r) = \{x : f_0(x) + \psi(x, r) \leqslant f_0(x_*)\}.$$

Ils sont compacts par hypothèse et

$$\Omega_m(r_2) \subset \Omega_m(r_1), \quad r_1 < r_2,$$

vu la croissance monotone de $\psi(x, r)$ avec r. Soit maintenant r_k, $k \to \infty$, une suite croissante de r et $r_k \to +\infty$. On a

$$\Omega_m(r_k) \subset \Omega_m(r_1).$$

Puisque $x(r) \in \Omega_m(r)$, tous les points $x(r_k)$ sont éléments de l'ensemble compact $\Omega_m(r_1)$. On dit donc sans restreindre la généralité que la suite $x(r_k)$ converge vers un point x_0.

Montrons que $x_0 \in \Omega$ et $f_0(x_0) = m$. En effet, si $x_0 \notin \Omega$, alors $\psi(x(r_k), r_k) \to +\infty$ et donc $F(x(r_k), r_k) \to +\infty$ car

$$f_0(x(r_k)) \geqslant \min_{x \in \Omega_m(r_1)} f_0(x).$$

Or, cela contredit $F(x(r_k), r_k) = m(r_k) \leqslant m$. Ainsi, $x_0 \in \Omega$.

Ensuite,

$$m(r_k) = F(x(r_k), r_k) = f_0(x(r_k)) + \psi(x(r_k), r_k) \leqslant$$
$$\leqslant f_0(x_0) + \psi(x_0, r_k) = f_0(x_0).$$

D'où

$$\lim_{k\to\infty} m(r_k) = \lim_{k\to\infty} (f_0(x(r_k)) + \psi(x(r_k), r_k)) \leqslant f_0(x_0). \tag{8.5}$$

Mais $f_0(x(r_k)) \to f_0(x_0)$, donc

$$\lim_{k\to\infty} \psi(x(r_k), r_k) \leqslant f_0(x_0) - \lim_{k\to\infty} f_0(x(r_k)) = 0.$$

Comme $\psi(x, r) \geqslant 0$, il s'ensuit

$$\lim_{k\to\infty} \psi(x(r_k), r_k) = 0.$$

Ainsi,

$$\lim_{k\to\infty} m(r_k) = \lim_{k\to\infty} f_0(x(r_k)) + \lim_{k\to\infty} \psi(x(r_k), r_k) = f_0(x_0) \geqslant m.$$

D'autre part, $m(r_k) \leqslant m$, ce qui fait que $\lim\limits_{k\to\infty} m(r_k) = f_0(x_0) \leqslant \leqslant m$. Comparons les deux dernières inégalités, il vient $\lim\limits_{k\to\infty} m(r_k) = = f_0(x_0) = m$, ce qui achève la démonstration du théorème.

Le théorème 8.1 montre qu'en ramenant le problème (8.1) à la recherche du minimum de $F(x, r)$ pour r importants on s'approche de la solution de celui-là. Evaluons cette convergence dans un cas non convexe en renvoyant le lecteur, pour le problème de programmation convexe, au numéro suivant.

THÉORÈME 8.2. *Etant donné les conditions du théorème* 8.1 *et des fonctions* $f_i(x)$, $i = 0, 1, \ldots, m$, *continûment dérivables, si*

1) *le problème* (8.1) *admet une solution unique*;

2) *la fonction* $\psi(x, r)$ *est prise sous forme* (8.3) *et le minimum de* $F(x, r)$ *est réalisé en un seul point* $x(r)$ *pour des* r *grands*;

3) *en* x_*, *solution du problème* (8.1), *les gradients* $f_i'(x_*)$, $i \in \in \mathcal{J}_0(x_*)$, *sont linéairement indépendants et*

$$\mathcal{J}_0(x_*) = \{i\colon f_i(x_*) = 0,\ i = 1, \ldots, m\},$$

alors $\lim\limits_{r\to\infty} x(r) = x_*$ *et*

$$\lim_{r\to\infty} r\varphi_1(f_i(x(r))) = \frac{1}{2} u^i, \quad i = 1, \ldots, m,$$

où u^i *sont les multiplicateurs de Lagrange de* (8.1).

Remarque. Rappelons que, selon le théorème 4.1 (§ 4, ch. I), le point x_* remplit les conditions nécessaires de minimum sous la forme

$$\begin{gathered} f_0'(x_*) + \sum_{i=1}^{m} u^i f_i'(x_*) = 0, \\ u^i \geqslant 0, \quad u^i f_i(x_*) = 0, \quad i = 1, \ldots, m. \end{gathered} \tag{8.6}$$

Démonstration. Commençons par montrer la convergence $x(r) \to x_*$. Procédons par l'absurde. Il existe alors une suite $r_k \to \infty$ telle que $\| x(r_k) - x_* \| \geqslant \delta_0 > 0$. Puisqu'en démontrant le théorème précédent nous avons montré l'appartenance $x(r_k) \in \Omega_m(r_1)$ et vu la compacité de ce dernier ensemble, on admet, sans nuire en rien à la généralité (on prend une sous-suite si le besoin en est), que $x(r_k) \to x_{**}$, et il est clair que $\| x_{**} - x_* \| \geqslant \delta_0 > 0$. Il résulte cependant du théorème 8.1 que x_{**} est solution du problème (8.1). Nous en avons donc obtenu deux solutions différentes, ce qui est contraire à l'hypothèse.

Ainsi, $x(r) \to x_*$. Démontrons la seconde affirmation du théorème. Comme $x(r)$ réalise le minimum de $F(x, r)$, le gradient de la fonction

$$F(x, r) = f_0(x) + r \sum_{i=1}^{m} \varphi_0(f_i(x))$$

doit être nul en ce point. Après de simples calculs on aboutit à l'égalité

$$F'(x(r), r) = f_0'(x(r)) + \sum_{i=1}^{m} (2r\varphi_1(f_i(x(r)))) f_i'(x(r)) = 0,$$

ou, avec les notations

$$\begin{gathered} u^i(r) = 2r\varphi_1(f_i(x(r))), \\ f_0'(x(r)) + \sum_{i=1}^{m} u^i(r) f_i'(x(r)) = 0. \end{gathered} \tag{8.7}$$

Notons à présent qu'étant donné $x(r) \to x_*$, on a $f_i(x(r)) < 0$ pour $i \notin \mathcal{J}_0(x_*)$ car $f_i(x_*) < 0$ pour $i \notin \mathcal{J}_0(x_*)$. Aussi, pour r grands

$$u^i(r) = 2r\varphi_1(f_i(x(r))) = 0, \quad i \notin \mathcal{J}_0(x_*).$$

Mais $u^i f_i(x_*) = 0$ et on a donc $u^i = 0$ quand $i \notin \mathcal{J}_0(x_*)$, ce qui, compte tenu de l'expression de $u^i(r)$, démontre l'affirmation du théorème pour i n'appartenant pas à $\mathcal{J}_0(x_*)$.

Ce que nous venons de dire permet de récrire (8.7) et (8.6) :

$$\begin{gathered} f_0'(x(r)) + \sum_{i \in \mathcal{J}_0(x_*)} u^i(r) f_i'(x(r)) = 0, \\ f_0'(x_*) + \sum_{i \in \mathcal{J}_0(x_*)} u^i f_i'(x_*) = 0. \end{gathered} \tag{8.8}$$

Si l'on prend en considération la convergence $x(r) \to x_*$, la continuité par rapport à x des $f_i'(x_*)$ et l'indépendance linéaire des $f_i'(x_*)$, $i \in \mathcal{J}_0(x_*)$, on déduit sans peine de (8.8) $u^i(r) \to u^i$, ce qui achève, compte tenu de l'expression de $u^i(r)$, la démonstration du théorème.

Remarque. Le théorème 8.2 entraîne que si $u^i > 0$, $f_i(x(r)) > 0$ pour r grands et $f_i(x(r))$ converge vers zéro comme $u^i r^{-1}$. Ainsi, une solution approchée viole toujours la contrainte $f_i(x) \leqslant 0$ si $u^i > 0$.

2. Programmation convexe. S'agissant du problème de programmation convexe, la convergence de $x(r)$ vers la solution cherchée x_* s'évalue plus exactement.

Théorème 8.3. *Si toutes les fonctions $f_i(x)$, $i = 0, 1, \ldots, m$, sont convexes, une fonction $\psi(x, r)$ de la forme* (8.3) *remplit les conditions du théorème* 8.1 *et on a en un point x_*, solution du problème* 8.1, *les conditions nécessaires sous forme du théorème de Kuhn et Tucker, i.e. il existe des nombres $u^i \geqslant 0$ tels que*

$$\begin{gathered} f_0(x_*) \leqslant \sum_{i=1}^{m} u^i f_i(x) + f_0(x), \quad \forall x, \\ u^i f_i(x_*) = 0, \quad i = 1, \ldots, m, \end{gathered} \tag{8.9}$$

on a alors

$$f_i(x(r)) \leqslant \frac{\bar{u}}{r} \quad si \quad f_i(x(r)) \geqslant 0, \tag{8.10}$$

$$f_0(x(r)) \geqslant f_0(x_*) - \frac{5}{4} \frac{\bar{u}^2}{r}, \tag{8.11}$$

où

$$\bar{u} = \sqrt{\sum_{i=1}^{m} (u^i)^2}.$$

Démonstration. Introduisons la notation $\tilde{\mathcal{J}}_0(x) = \{i\colon f_i(x) \geqslant 0, i = 1, \ldots, m\}$. Comme

$$F(x(r), r) = f_0(x(r)) + r \sum_{i=1}^{m} \varphi_0(f_i(x(r))) \leqslant f_0(x_*),$$

il découle de (8.9)

$$f_0(x(r))+r\sum_{i=1}^{m}\varphi_0(f_i(x(r)))\leqslant f_0(x(r))+\sum_{i=1}^{m}u^i f_i(x(r)),$$

ou

$$r\sum_{i=1}^{m}\varphi_0(f_i(x(r)))\leqslant\sum_{i=1}^{m}u^i f_i(x(r)).$$

Mais pour $i\notin\widetilde{\mathcal{J}}_0(x(r))$

$$\varphi_0(f_i(x(r)))=0,\quad f_i(x(r))<0,$$

et $\varphi_0(f_i(x(r)))=f_i^2(x(r))$ pour $i\in\widetilde{\mathcal{J}}_0(x(r))$. On renforce donc l'inégalité obtenue :

$$r\sum_{i\in\widetilde{\mathcal{J}}_0(x(r))}f_i^2(x(r))\leqslant\sum_{i\in\widetilde{\mathcal{J}}_0(x(r))}u^i f_i(x(r))\leqslant\bar{u}\sqrt{\sum_{i\in\widetilde{\mathcal{J}}_0(x(r))}f_i^2(x(r))}.$$

Le dernier résultat est obtenu moyennant l'inégalité connue de Cauchy-Bouniakovski.

Ainsi,

$$f_i(x(r))\leqslant\sqrt{\sum_{i\in\widetilde{\mathcal{J}}_0(x(r))}f_i^2(x(r))}\leqslant\frac{\bar{u}}{r},\tag{8.12}$$

d'où (8.10).

Quels que soient x, on a

$$f_0(x_*)\leqslant f_0(x)+\sum_{i=1}^{m}u^i f_i(x)\leqslant f_0(x)+\sum_{i\in\widetilde{\mathcal{J}}_0(x)}u^i f_i(x)=$$

$$=f_0(x)+r\sum_{i\in\widetilde{\mathcal{J}}_0(x)}f_i^2(x)-\sum_{i\in\widetilde{\mathcal{J}}_0(x)}\left(\sqrt{r}f_i(x)-\frac{u^i}{2\sqrt{r}}\right)^2+$$

$$+\sum_{i\in\widetilde{\mathcal{J}}_0(x)}\frac{(u^i)^2}{4r}\leqslant f_0(x)+r\sum_{i=1}^{m}\varphi_0(f_i(x))+\frac{\bar{u}^2}{4r}=F(x,r)+\frac{\bar{u}^2}{4r}.$$

Ceci étant, on tient compte de l'égalité

$$\sum_{i=1}^{m}\varphi_0(f_i(x))=\sum_{i\in\widetilde{\mathcal{J}}_0(x)}f_i^2(x),\tag{8.13}$$

qu'on a par définition de $\varphi_0(x)$ et $\widetilde{\mathcal{J}}_0(x)$. Ainsi,

$$f_0(x_*)\leqslant f_0(x(r))+r\sum_{i=1}^{m}\varphi_0(f_i(x(r)))+\frac{\bar{u}^2}{4r}.$$

Mais (8.12) et (8.13) entraînent

$$\sum_{i=1}^{m} \varphi_0 (f_i (x (r))) \leqslant \frac{\overline{u}^2}{r^2}.$$

C'est pourquoi

$$f_0 (x (r)) \geqslant f_0 (x_*) - \frac{5}{4} \frac{\overline{u}^2}{r},$$

c.q.f.d.

3. Aspects numériques. Agissant comme ci-dessus on ramène le problème (8.1) à la minimisation de la fonction $F(x, r)$. La solution approchée s'obtient maintenant par l'une des méthodes du ch. II. Certaines circonstances sont cependant à retenir. Dans le cas des $f_i(x)$ non convexes, $F(x, r)$ n'est pas non plus convexe en x. Cette fonction peut présenter donc des minima locaux. Or, dans tout l'exposé précédent il s'agissait de trouver le minimum global $x(r)$.

Nous avons vu que dans le cas de fonctions non convexes, toutes les techniques du chapitre II permettent de trouver un minimum local de $x(r)$ pour une mauvaise approximation initiale. Cela altère la convergence et nuit beaucoup à la qualité de la méthode de pénalisation en ce qui concerne les problèmes non convexes.

S'agissant d'un problème de programmation convexe avec pour $\psi(x, r)$ la fonction (8.3), $F(x, r)$ est également convexe, fait qui s'établit sans peine. On n'a donc plus à surmonter la difficulté mentionnée. Mais un autre obstacle survient. Le fait est qu'une bonne approximation s'obtient en prenant un r suffisamment grand (cela résulte des estimations ci-dessus). Toutes les dérivées de $F(x, r)$ par rapport à x sont alors grandes elles aussi parce que proportionnelles à r. Lorsque nous avons analysé les méthodes à convergence superlinéaire exposées dans le ch. II, nous avons établi que la dimension du voisinage dans lequel la convergence devient superlinéaire est inversement proportionnelle à la constante de la condition de Lipschitz vérifiée par les dérivées secondes, i.e. ce voisinage est également petit dans le cas considéré, et même une méthode qui converge théoriquement bien à la limite peut s'avérer inefficace. Il y a plus. Comme la fonction $\varphi_0(t)$ ne possède pas pour $t = 0$ de dérivée seconde, $F(x, r)$ calculée moyennant $\psi(x, r)$ de la formule (8.3) en est également dépourvue aux points x en lesquels $f_i(x) = 0$ pour un i. Mais si la solution x_* appartient à la frontière du domaine, c'est ce cas qui a justement lieu. D'autre part, toutes les méthodes rapidement convergentes exigent que la fonction à minimiser possède des dérivées secondes au moins dans un certain voisinage du point cherché.

Les difficultés énumérées se répercutent de règle sur les calculs réels, ce qui détériore la qualité de la méthode.

4. Méthode de Fiacco et McCormick. Cette méthode ne s'applique en fait elle aussi qu'à des problèmes de programmation convexe. Son idée est proche de celle de la méthode de pénalisation mais la différence est qu'on s'approche de la solution de l'intérieur du domaine et non pas de l'extérieur.

Revenons au problème (8.1) en supposant que toutes les $f_i(x)$ sont convexes et qu'il existe un point $\bar{x}$ tel que $f_i(\bar{x}) < 0$, $i = 1, \ldots, m$, de sorte que l'ensemble admissible Ω soit d'intérieur non vide. Formons la fonction

$$P(x, r) = f_0(x) - \sum_{i=1}^{m} \frac{r}{f_i(x)}, \qquad r > 0,$$

définie à l'intérieur de Ω. On vérifie aisément que $P(x, r)$ est convexe par rapport à x à l'intérieur de Ω. En désignant par $x(r)$ le point de minimum de $P(x, r)$ dans Ω on montre, dans des hypothèses assez générales analogues à celles des théorèmes 8.1 et 8.2, que

$$\lim_{r \to +0} x(r) = x_*,$$

$$\lim_{r \to 0} \frac{r}{f_i^2(x(r))} = u^i, \quad i = 1, \ldots, m.$$

Ainsi, la résolution approchée de (8.1) s'est encore réduite au problème de la recherche du minimum de $P(x, r)$ sans contraintes.

En ce qui concerne les particularités de ce problème auxiliaire, on n'a qu'à répéter ce que nous avons dit au n° 3 au sujet de la méthode de pénalisation. Un exemple simple suffit pour les illustrer et montrer pourquoi même des méthodes efficaces de minimisation de $F(x, r)$ et $P(x, r)$ peuvent ne pas garantir une convergence rapide.

Soit $f_0(x) = -x$, $f_1(x) = x$, $x \in E^1$, i.e. on demande de trouver le minimum de $-x$ avec la contrainte $x \leqslant 0$. Pour la solution évidente $x_* = 0$, on a

$$P(x, r) = -x - \frac{r}{x}.$$

Egalons à zéro la dérivée de $P(x, r)$ par rapport à x, il vient

$$P'(x, r) = -1 + \frac{r}{x^2} = 0, \tag{8.14}$$

d'où $x(r) = -\sqrt{r}$. Appliquons à (8.14) une méthode à convergence quadratique, à savoir la méthode de Newton, i.e. les approximations seront générées par la formule

$$x_{k+1} = x_k - \frac{P'(x_k, r)}{P''(x_k, r)}.$$

Portons-y les expressions de $P'(x, r)$ et $P''(x, r)$ et effectuons de simples transformations. Nous obtenons

$$v_{k+1} = \frac{2\sqrt{r} - x_k}{2r} v_k^2, \quad v_k = x_k + \sqrt{r}. \tag{8.15}$$

Il ressort de cette formule que l'écart de x_k par rapport à la solution $x(r) = -\sqrt{r}$ tend monotonement vers zéro pour les seuls points initiaux en lesquels

$$\frac{2\sqrt{r} - x_k}{2r} |v_k| < 1.$$

Comme $x_k < 0$ (l'approximation est cherchée dans le domaine $x < 0$), on a

$$\frac{2\sqrt{r}}{2r} |v_k| < 1, \quad |v_k| < \sqrt{r}.$$

Ainsi, la dernière formule montre que la convergence quadratique de la méthode de Newton n'est garantie que dans un domaine où x_k s'écarte de la solution d'au plus $\sqrt{r}$, i.e. le domaine de convergence de la méthode tend vers zéro avec la diminution de r, et tel que sa dimension soit de l'ordre de grandeur de l'écart entre $x(r)$ et la vraie solution x_* du problème initial. L'effort de calcul principal est donc fait pour tomber dans le domaine de convergence de la méthode de Newton, et là où celle-ci converge bien on n'en a plus besoin du moment qu'on dispose déjà d'une approximation dont l'écart de x_* vaut celui entre x_* et $x(r)$.

§ 9. Méthodes de projection avec reconstitution des liaisons

1. Schéma de construction des méthodes. Considérons le problème de minimisation de la fonction $f_0(x)$ dans les conditions

$$f_i(x) = 0, \quad i = 1, \ldots, m, \quad m < n. \tag{9.1}$$

Notons $g = (f_1, \ldots, f_m)$, $S_g = \{x : g(x) = 0\}$. Supposons que toutes les fonctions $f_0(x)$, $f_1(x)$, ..., $f_m(x)$ sont continûment dérivables et que S_g est une variété régulière ($(n-m)$-dimensionnelle), i.e. en tout point $x \in S_g$ le rang de la matrice $g'(x)$ vaut m ($g'(x) = \left\{\frac{\partial f_i(x)}{\partial x^j}\right\}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, i est l'indice de la ligne). Par conséquent, on construit en tout point $\tilde{x} \in S_g$ un hyperplan tangent à S_g:

$$g'(\tilde{x})(x - \tilde{x}) = 0. \tag{9.2}$$

Cet hyperplan (i.e. l'ensemble de points vérifiant l'équation (9.2)) sera noté $T(\tilde{x})$.

En construisant les processus itératifs de résolution du problème formulé on s'inspire des considérations suivantes.

Soit x_0 un point de S_g tel que le gradient $f_0'(x_0)$ ne soit pas orthogonal à l'hyperplan $T(x_0)$ (i.e. le point x_0 ne remplit pas une condition nécessaire d'extrémum de $f_0(x)$ sur la variété S_g). Il existe alors dans $T(x_0)$ une infinité de directions de descente de $f_0(x)$ (i.e. une infinité de directions $x - x_0 \in T(x_0)$ telles que $(f_0'(x_0), x - x_0) < 0$). Supposons définie l'une de ces directions, disons $v_0 = \bar{x}_0 - x_0$, et construit un point $\hat{x}_0(\alpha) = x_0 + \alpha v_0$ tel que $f_0(\hat{x}_0(\alpha)) < f_0(x_0)$. Le point $\hat{x}_0$ ne satisfait plus aux équations de liaison (9.1). Si la valeur du paramètre α est suffisamment petite (i.e. si la quantité $\| x_0 - \hat{x}_0(\alpha) \|$ est faible) on construit cependant de diverses façons, à l'aide de $\hat{x}_0(\alpha)$, un point $x_1 \in S_g$ tel que

$$f_0(x_1) < f_0(x_0). \tag{9.3}$$

En effet, on peut choisir sur la variété régulière S_g un point $x_1(\alpha)$ (qui n'est pas d'ailleurs unique) de sorte à vérifier la condition

$$x_1(\alpha) - x_0 = \hat{x}_0(\alpha) - x_0 + \omega_0(\alpha),$$

où

$$\| \omega_0(\alpha) \| \equiv \| x_1(\alpha) - \hat{x}_0(\alpha) \| = o(\| \hat{x}_0(\alpha) - x_0 \|). \tag{9.4}$$

(Pour la démonstration rigoureuse utiliser le théorème sur l'application qui transforme l'un dans l'autre le voisinage du point x_0 dans la variété S_g et celui dans la variété tangente $T(x_0)$, théorème qui est valable dans l'espace E^n; voir L. Ljusternik, W. Sobolew [86]).

Lorsque (9.4) a lieu, on a par suite de la dérivabilité de $f_0(x)$:

$$\begin{aligned} f_0(x_1) &= f_0(\hat{x}_0) + (f_0'(\hat{x}_0),\ x_1 - \hat{x}_0) + o(\| x_1 - \hat{x}_0 \|) = \\ &= f_0(x_0) + (f_0'(x_0),\ \hat{x}_0 - x_0) + o(\| \hat{x}_0 - x_0 \|) + \\ &+ (f_0'(\hat{x}_0),\ x_1 - \hat{x}_0) + o(\| x_1 - \hat{x}_0 \|) = f_0(x_0) + \\ &+ (f_0'(x_0),\ \hat{x}_0 - x_0) + o_1(\| \hat{x}_0 - x_0 \|). \end{aligned}$$

D'où l'inégalité (9.3) si le paramètre α est suffisamment petit.

En construisant un point $x_1 \in S_g$ vérifiant la condition (9.3), nous avons en fait terminé l'itération d'un processus de descente qui génère les approximations successives de la solution. La k-ième itération du processus du type décrit comprend donc les étapes suivantes.

1. On définit une direction de descente $v_k = \bar{x}_k - x_k$ de la fonction $f_0(x)$ dans l'hyperplan tangent $T(x_k)$.

2. On fait un pas de longueur déterminée dans la direction v_k: $\hat{x}_k(\alpha) = x_k + \alpha v_k$ (de façon à avoir $f_0(\hat{x}_k) < f_0(x_k)$).

3. Moyennant le point $\hat{x}_k(\alpha)$ on définit un point $x_{k+1} \in S_g$ en lequel a lieu la condition $f_0(x_{k+1}) < f_0(x_k)$.

L'exposé précédent montre qu'on se déplace du point x_k suivant différentes directions de descente dans le plan $T(x_k)$. Le choix de la quantité α_k et la partie finale de l'itération (la construction du point x_{k+1}) ne se définissent pas eux non plus de façon unique. A force d'effectuer différemment chaque étape de l'itération on construit toute une classe de processus de descente du type décrit.

Voyons certaines façons dont on peut choisir v_k. On peut prendre pour v_k la projection de l'opposé du gradient $-f_0'(x_k)$ sur $T(x_k)$. Cette construction équivaut à la résolution du problème de minimisation de la fonction

$$F_k(x) = (f_0'(x_k),\ x - x_k) = \frac{1}{2} \| x - x_k \|^2 \tag{9.5}$$

dans la condition $x \in T(x_k)$. En utilisant la méthode des multiplicateurs de Lagrange on établit que

$$v_k = -(I - g'^* (g'g'^*)^{-1} g') f_0'(x_k), \tag{9.6}$$

avec $g' = g'(x_k)$.

On construit des méthodes plus efficaces en choisissant en qualité de v_k un vecteur minimisant la fonction

$$F_k(x) = (f_0'(x_k),\ x - x_k) + \frac{1}{2} (f_0''(x_k)(x - x_k),\ x - x_k) \tag{9.7}$$

sur le plan $T(x_k)$ (ce vecteur existe si F_k est convexe). Puisque la direction de déplacement est ici construite en fait à l'aide de l'approximation quadratique de la fonction à minimiser, les méthodes dans lesquelles v_k est formé de façon décrite seront dites *du second ordre*.

Considérons le procédé de reconstitution des liaisons (troisième étape de l'itération) dont nous nous servirons dans la suite.

Supposons que le système d'équations (9.1) définit dans un voisinage de tout point $x \in S_g$ une fonction $y = y(z)$ avec y un m-vecteur des axes de coordonnées et z un $(n - m)$-vecteur. On pose sans restreindre la généralité: $y = (x^1, \ldots, x^m)$, $z = (x^{m+1}, \ldots, x^n)$. Selon le théorème de fonctions implicites, pour que la fonction $y(z)$ et ses dérivées existent, il faut qu'en tout point $x \in S_g$ le déterminant

$$g_y(x)| = \left| \left\{ \frac{\partial f_i}{\partial x^j} \right\} \right| \neq 0, \quad i, j = 1, \ldots, m. \tag{9.8}$$

Dans ce cas le point $x_{k+1} = (z_{k+1}, y_{k+1})$ est généré par les formules

$$z_{k+1} = z_k + \alpha_k p_k, \quad y_{k+1} = y(z_{k+1}), \tag{9.9}$$

où $p_k = \bar{z}_k - z_k$ est la partie correspondante du vecteur v_k. La formation de la suite (9.9) n'exige pas qu'on explicite la fonction

$y(z)$; il suffit de savoir calculer sa valeur (i.e. résoudre le système (9.1)) pour z fixe.

A part qu'elle permet la mise en œuvre d'un procédé de reconstitution des liaisons, la génération de la suite (9.9) s'interprète comme processus itératif de minimisation de la fonction $\varphi(z) = f_0(z, y(z))$. Il est évident que dans la condition (9.8) la recherche du minimum de $\varphi(z)$ équivaut à la résolution du problème de départ. Le vecteur p_k (une direction de descente de la fonction $\varphi(z)$) est alors considéré comme solution du problème de minimisation de $\psi_k(z) = F_k(z, y_1(z))$, où $F_k(z, y)$ se définit par l'une des formules (9.5) ou (9.7) et la fonction vectorielle $y_1(z)$ par l'équation de liaison linéarisée (i.e. par l'équation donnant le plan tangent $T(x_k)$) :

$$g_y(x_k)(y - y_k) + g_z(x_k)(z - z_k) = 0.$$

D'où

$$y_1(z) = y_k - g_y^{-1}(x_k)\, g_z(x_k)(z - z_k).$$

Le fait que le vecteur p_k défini de manières décrites est une direction de descente de $\varphi(z)$ découle de $\psi_k'(z_k) = \varphi'(z_k)$, où

$$\varphi'(z_k) = f_{0z}(x_k) + y'^*(z_k)\, f_{0y}(x_k),$$

$$f_{0z} = \left(\frac{\partial f_0}{\partial x^{m+1}}, \ldots, \frac{\partial f_0}{\partial x^n}\right), \quad f_{0y} = \left(\frac{\partial f_0}{\partial x^1}, \ldots, \frac{\partial f_0}{\partial x^m}\right), \tag{9.10}$$

$$y'(z_k) = -g_y^{-1}(x_k)\, g_z(x_k).$$

Puisque le processus du type (9.9) peut s'interpréter comme méthode de minimisation de $\varphi(z)$, on conçoit qu'on peut choisir pour p_k le vecteur $-\varphi'(z_k)$ auquel cas la suite (9.9) représente une méthode du gradient pour la recherche du minimum de $\varphi(z)$.

Notons que le vecteur p_k réalisant le minimum de la fonction $F_k(z, y_1(z))$, avec $F_k(z, y)$ défini par (9.5), se calcule par la formule

$$p_k = -(I + y'^*(z_k)\, y'(z_k))^{-1}\, \varphi'(z_k). \tag{9.11}$$

La suite (9.9) avec p_k défini par (9.11) constitue donc elle aussi une méthode du type gradient destinée à rechercher le minimum de $\varphi(z)$. *Les méthodes du type gradient seront dites du premier ordre.*

Si l'on recherche le minimum de $\varphi(z)$ dans les conditions nécessaires, on recourt en principe à la méthode de Newton et à ses variantes. Notons cependant que le calcul de la dérivée seconde $\varphi''(z)$ s'avère en général fort laborieux parce qu'exigeant la connaissance de la dérivée seconde de la fonction vectorielle $y(z)$, i.e. en fait des dérivées secondes des fonctions $f_1(x), \ldots, f_m(x)$.

Supposons à présent qu'on a, au lieu de (9.8), une condition moins sévère : en tout point $x \in S_g$ au moins un déterminant d'ordre

m est différent de zéro :

$$\left| \left\{ \frac{\partial f_i}{\partial x^j} \right\} \right| \neq 0, \tag{9.12}$$

$$j = j_1, j_2, \ldots, j_m, \quad j_i \in [1, \ldots, n], \quad i = 1, \ldots, m.$$

Les fonctions f_i sont astreintes à des conditions plus faibles : les différents déterminants sont non nuls en des points divers de l'ensemble S_g. Dans ce cas, les coordonnées d'un point $x \in S_g$, qui constituent le vecteur z et la fonction vectorielle $y(z)$, diffèrent en général selon le point de la variété S_g : $z = (x^{j_{m+1}}, \ldots, x^{j_n})$, $y = (x^{j_1}, \ldots, x^{j_m})$. Ceci étant, on reconstitue toujours les liaisons par les formules (9.9). Chaque pas du processus (9.9) s'interprète alors comme celui du processus de minimisation d'une fonction $\varphi(x^{j_{m+1}}, \ldots, x^{j_n})$ pour laquelle le vecteur p_k correspondant est une direction de descente.

Nous nous occuperons plus loin des méthodes du type (9.9) et nous noterons y toute fonction vectorielle $(x^{j_1}, \ldots, x^{j_m})$ et z le vecteur variables indépendantes (comme on l'a fait sous la condition (9.8)). Tout déterminant $|\{\partial f_i/\partial x^j\}|$ d'ordre m sera donc désigné par $|g_y|$ et la fonction $f_0(z, y(z))$ par $\varphi(z)$. La valeur absolue de la fonction $|g_y(x)|$ sera notée $|g_y(x)|_A$.

Les deux numéros suivants seront consacrés aux propriétés des méthodes du premier et du second ordre. Le n° 4 traitera des méthodes utilisant des directions duales ou conjuguées pour minimiser $\varphi(z)$ (ou des algorithmes s'appuyant sur des méthodes de ce type). Du point de vue de la pratique, ce sont ces algorithmes qui offrent le plus d'intérêt.

2. Méthodes du premier ordre. Etudions les méthodes utilisant la linéarisation de la fonction $f_0(x)$ et des liaisons f_i, $i = 1, \ldots, m$.

Considérons un algorithme dont chaque pas est celui d'une méthode du gradient destinée à minimiser une fonction $\varphi(z)$:

$$z_{k+1} = z_k - \alpha_k \varphi'(z_k), \; y_{k+1} = y(z_{k+1}), \tag{9.13}$$

où z_k est le vecteur correspondant au déterminant $|\bar{g}_y(x)|$ de plus grande valeur absolue au point $x_k \in S_g$ parmi les déterminants $|g_y|$, le gradient $\varphi'(z_k)$ s'obtient moyennant la formule (9.10) et le paramètre α_k se définit par l'un des procédés décrits au § 1, ch. II. Nous prendrons pour α_k la valeur maximale du paramètre obtenue par fractionnement à partir d'une constante positive satisfaisant à l'inégalité

$$f_0(z, y(z)) - f_0(z_k, y_k) \leqslant -\varepsilon\alpha \, \| \varphi'(z_k) \|^2, \quad 0 < \varepsilon < 1, \tag{9.14}$$

avec $z = z_k - \alpha\varphi'(z_k)$ (l'analogue du procédé de choix de α_k à partir de la condition (II.1.2.)).

THÉORÈME 9.1. *Si les fonctions $f_0(x)$ et $f_i(x)$, $i = 1, \ldots, m$, sont deux fois continûment dérivables, les fonctions f_i vérifiant la condition* (9.12) *et l'ensemble $S = S_g \cap S_0$ ($S_0 = \{x: f_0(x) \leqslant f_0(x_0)\}$) étant borné pour un choix arbitraire du point x_0, alors $f_0(x_{k+1}) \leqslant \leqslant f_0(x_k)$ sur la suite* (9.13) *et $\| \varphi'(z_k) \| \to 0$ pour $k \to \infty$.*

DÉMONSTRATION. La condition (9.12) entraîne la possibilité de construire la suite (9.13): pour les valeurs suffisamment faibles du paramètre α le point z_{k+1} tombe dans le voisinage du point z_k, domaine de définition de la fonction $y(z)$. En vertu des conditions du théorème la fonction $\varphi(z) = f(z, y(z))$ y est deux fois continûment dérivable. On a donc l'estimation

$$\varphi(z_{k+1}) - \varphi(z_k) \leqslant \alpha_k \| \varphi'(z_k) \|^2 \left(-1 + \frac{\alpha_k}{2} \| \varphi''(z_c) \| \right), \qquad (9.15)$$

avec $\varphi''(z_c) = \varphi''(z_k + \theta(z_{k+1} - z_k))$, $\theta \in [0, 1]$, d'où l'inégalité (9.14) pour les valeurs suffisamment petites de α_k. Cela signifie que la fonction $f_0(x)$ diminue de façon monotone sur les termes de la suite (9.13).

Démontrons que $\| \varphi'(z_k) \| \to 0$. Une fonction continue $|\bar{g}_y(x)|_A$ atteint sa valeur minimale γ sur un ensemble fermé borné S (théorème de Weierstrass), et, en vertu de (9.12), $\gamma > 0$ (la fonction $|\bar{g}_y(x)|_A$ est continue en tant que la plus grande des fonctions continues $|g_y(x)|_A$). Il existe un nombre fini de déterminants fonctionnels $|g_y|$ distincts et, vu la dérivabilité des fonctions f_i, ils sont tous des fonctions uniformément continues sur S. On trouve donc pour toute constante $0 < \gamma_1 < \gamma$ une constante $\rho > 0$ telle qu'en tout point de l'ensemble S appartenant à une sphère S_ϑ de rayons ρ et de centre dans un point quelconque $\vartheta \in S$, la valeur absolue du déterminant $|g_y(x)|$ prenant la valeur $|\bar{g}_y(\vartheta)|$ en ϑ soit au moins γ_1. L'ensemble S étant borné et les dérivées partielles premières et secondes des f_i continues, ces dérivées sont bornées (par une constante M) quel que soit $\vartheta \in S$ choisi dans la sphère S_ϑ de rayon ρ. Ceci étant, on dit, en conformité avec les théorèmes de fonctions implicites, que dans un parallélépipède

$$[\vartheta^i \pm \delta\vartheta^i], \; i = 1, \ldots, n,$$

appartenant à la sphère S_ϑ le système (9.1) définit au moins une fonction vectorielle $y(z)$ deux fois continûment dérivable et que dans ce parallélépipède une telle fonction a ses dérivées bornées:

$$\| y'(z) \| \leqslant N_1, \quad \| y''(z) \| \leqslant N_1. \qquad (9.16)$$

Vu cette circonstance et la propriété analogue des dérivées premières et secondes de la fonction $f_0(x)$ sur S, les dérivées de $\varphi(z)$ sont également bornées dans le parallélépipède $[\vartheta^i \pm \delta\vartheta^i]$: $\| \varphi'(z) \| \leqslant N_2$, $\| \varphi''(z) \| \leqslant N_2$.

On se convainc maintenant qu'il existe une constante $\delta > 0$ telle que si $\alpha_k \leqslant \delta$, le point x_{k+1} se trouve dans une sphère S_{x_k} de rayon ρ. En effet, supposons que $\| x_{k+1} - x_k \|^2 = \| z_{k+1} - z_k \|^2 + \| y_{k+1} - y_k \|^2 = \rho^2$. Ces égalités signifient l'appartenance de x_{k+1} à S_{x_k} et ce point est donc élément d'un parallélépipède $[x_k^i \pm \delta x_k^i]$, $i = 1, \ldots, n$, dans lequel les dérivées de la fonction $y(z)$ s'évaluent par (9.16). Par conséquent, $\| y_{k+1} - y_k \| \leqslant N_3 \| z_{k+1} - z_k \|$, et il résulte des égalités précédentes que $(N_3^2 + 1) \| z_{k+1} - z_k \|^2 = \alpha_k^2 N_4^2 \| \varphi'(z_k) \|^2 \geqslant \rho^2$, d'où

$$\alpha_k \geqslant \frac{\rho}{N_4 \| \varphi'(z_k) \|} \geqslant \frac{\rho}{N_4 N_2}.$$

Selon cette estimation l'égalité $\| x_{k+1} - x_k \| = \rho$ a lieu pour le paramètre $\alpha_k \geqslant \rho/(N_4 N_2)$, i.e. on choisit pour δ n'importe quelle constante au plus égale à $\rho/(N_4 N_2)$.

Utilisons l'inégalité (9.15) (et prenons en considération le caractère borné de $\varphi''(z)$) ; nous nous assurons sans peine que l'inégalité (9.14) a manifestement lieu pour $\alpha_k = \min \left\{ \delta, \frac{2(1-\varepsilon)}{N_2} \right\}$, ce qui veut dire, $f_0(x)$ étant minoré (sur l'ensemble S), qu'on a nécessairement $\| \varphi'(z_k) \| \to 0$ quand k tend vers l'infini. Le théorème est démontré.

Dans le cas général, la dernière condition signifie la convergence de la suite (9.13) (ou d'une suite partielle de (9.13)) vers le point x_* qui vérifie une condition nécessaire d'extrémum de $f_0(x)$ sur la variété S_g (en ce point le gradient $f_0'(x_*)$ est orthogonal à l'hyperplan tangent $g'(x_*)(x - x_*) = 0$; voir ch. I, § 4). Comme $f_0(x)$ est continue, elle admet un minimum sur S. Si la suite (9.13) converge vers la solution et la fonction $\varphi(z)$, qui est la fonction à minimiser du problème auxiliaire auquel se réduit le problème initial dans un voisinage du minimum, vérifie les conditions $m_0 \| v \|^2 \leqslant (\varphi''(z)\, v, v) \leqslant M \| v \|^2$ pour tout $v \in E^{n-m}$, la convergence est au moins aussi rapide que celle d'une progression géométrique — cela découle des résultats généraux sur la convergence des méthodes du gradient (théorème II.1.2). Arrêtons-nous sur certaines questions relatives à la mise en œuvre de la méthode.

Dans le théorème que nous venons de démontrer on a défini les vecteurs z_k et y_k à l'aide du déterminant $|\bar{g}_y(x_k)|$. On le trouve à chaque itération en calculant tous les déterminants $|g_y|$. Dans la pratique, on n'en a cependant pas besoin (si l'on s'est servi de $|\bar{g}_y|$ dans le théorème, c'est uniquement pour simplifier la démonstration). La convergence de la méthode ne change pas en choisissant z_k et y_k correspondant à tout déterminant $|g_y|$ dont la valeur absolue au point x_k est au moins égale à une constante positive arbitraire μ (la même pour tous les k). Dans les conditions du théorè-

me cette constante existe puisqu'il en est de même de la constante γ. Pour mettre en œuvre l'algorithme on doit donc choisir les vecteurs z et y correspondant à un même déterminant tant qu'en un point il ne devient pas en module inférieur à μ; c'est alors qu'on passe à d'autres z et y, i.e. on calcule un autre déterminant $|g_y|$. La constante μ est choisie quelconque. Il se peut qu'en un point x_k tous les déterminants $|g_y|$ soient de module inférieur à μ. On prendra alors une autre constante $\mu_1 < \mu$. Chaque partition du paramètre α nécessaire pour qu'on ait l'inégalité (9.14) demande qu'on recalcule la fonction $y(z)$ (pour obtenir $f_0(x) = f_0(z, y(z))$), i.e. on résout le système d'équations non linéaires (9.1) pour z fixe. Pour réduire le volume de calcul on trouve la valeur cherchée du paramètre en testant l'inégalité

$$f_0(z, y_1(z)) - f_0(z_k, y_k) \leqslant -\varepsilon\alpha \|\varphi'(z_k)\|^2, \quad 0 < \varepsilon < 1. \quad (9.17)$$

Dès que (9.17) a lieu, on vérifie l'inégalité (9.14) pour α obtenu; si cette dernière n'est pas satisfaite, on continue de fractionner α, et si elle l'est, on se contente de ce α ou on essaie de l'augmenter en testant toujours (9.14). Notons que

$$y(z) = y_k + y'(z_k)(z - z_k) + O(\|z - z_k\|^2) = y_1(z) + O(\|z - z_k\|^2).$$

Pour $z_{k+1} - z_k$ suffisamment petit, on a $f_0(z_{k+1}, y_1(z_{k+1})) \to$ $\to f_0(z_{k+1}, y(z_{k+1}))$, et donc (9.14) a lieu simultanément avec (9.17) i.e. on se passe de fractionnements complémentaires du pas.

Remarque. Dans le théorème 9.1 on peut affaiblir les conditions imposées à la régularité des fonctions $f_0(x)$ et $f_i(x)$, ce qui complique du même coup la démonstration.

Voyons en quelques mots une autre méthode du type (9.9) dans laquelle le vecteur p_k est choisi d'après la formule (9.11) (les vecteurs z_k, y_k se définissent comme dans la méthode précédente) et α_k est la valeur maximale du paramètre (qu'on obtient par fractionnement) pour laquelle on a l'inégalité

$$f_0(z, y(z)) - f_0(z_k, y_k) \leqslant \varepsilon\alpha(\varphi'(z_k), p_k), \quad z = z_k + \alpha p_k.$$

Le théorème 9.1 reste en vigueur pour cet algorithme, et la démonstration ne diffère que par certains détails (comme pour les théorèmes sur les propriétés des méthodes du type gradient et ceux relatifs à la méthode de la plus grande pente du § 1, ch. II).

Notons que le coût par itération d'un tel algorithme dépasse celui de la méthode (9.13).

3. Une méthode du second ordre. Supposons $f_0(x)$ fortement convexe. La fonction quadratique $F_k(x)$ (9.7) est alors strictement convexe de même que $\psi_k(z) = F_k(z, y_1(z))$ (étant donné la linéarité de $y_1(z)$). Plus précisément, en vertu de la convexité forte

de $f_0(x)$, toute fonction $\psi_k(z)$ vérifie, quel que soit le vecteur $v \in E^{n-m}$, les conditions

$$m_0 \| v \|^2 \leqslant (\psi_k'' v, v) \leqslant M_0 \| v \|^2, \quad m_0 > 0, \tag{9.18}$$

avec la matrice $\psi_k'' = f_{0zz} + y'^* f_{0yy} y' + 2y'^* f_{0zy}$ (toutes les dérivées se calculent au point x_k). Dans ce cas, le vecteur p_k minimisant $\psi_k(z)$ est obtenu par la formule

$$p_k = -(\psi_k'')^{-1} \psi_k'(z_k). \tag{9.19}$$

Dans la méthode du second ordre le point x_{k+1}, $k = 0, 1, \ldots$, est construit comme suit :

$$z_{k+1} = z_k - \alpha_k (\psi_k'')^{-1} \psi'(z_k), \quad y_{k+1} = y(z_{k+1}), \tag{9.20}$$

où les vecteurs z_k et y_k se définissent comme dans la méthode (9.13) et où l'on prend pour α_k la plus grande valeur (obtenue par fractionnement) du paramètre, qui vérifie l'inégalité

$$f_0(x) - f_0(x_k) \leqslant \varepsilon\alpha (\varphi'(z_k), p_k), \quad 0 < \varepsilon < \frac{1}{2}. \tag{9.21}$$

Ici $x = (z, y(z))$, $z = z_k + \alpha p_k$.

THÉORÈME 9.2. *Soit une fonction deux fois continûment dérivable* $f_0(x)$,

$$m \| \omega \|^2 \leqslant (f_0''(x)\omega, \omega) \leqslant M \| \omega \|^2, \quad m > 0,$$

pour tout vecteur $\omega \in E^n$ *et les fonctions* $f_i(x)$, $i = 1, \ldots, m$, *remplissant les exigences du théorème* 9.1. *Quel que soit le point* x_0 *choisi, les résultats du théorème* 9.1 *se conservent pour la méthode* (9.20).

Le schéma de démonstration est analogue à celui du théorème 9.1. Nous ne nous intéresserons donc qu'à ce que le choix différent du vecteur p_k y apporte du nouveau.

Etant donné la stricte convexité de $f_0(x)$, l'ensemble S_0 est borné. Par conséquent, l'ensemble $S = S_g \cap S_0$ est borné et fermé (vu la fermeture de S_0 et de S_g). On le prend en considération et on établit, comme dans le théorème 9.1, les estimations (9.16) et la borne dans le parallélépipède $[\vartheta^i \pm \delta\vartheta^i]$, $i = 1, \ldots, n$, des dérivées $\varphi'(z)$, $\varphi''(z)$.

Ensuite, on a $\| (\psi_k'')^{-1} \| \leqslant \frac{1}{m_0}$ en vertu de (9.18), donc $\| p_k \| = \| (\psi_k'')^{-1} \psi_k'(z_k) \| = \| (\psi_k'')^{-1} \varphi'(z_k) \| \leqslant N_2/m_0$ et, partant,

$$\alpha_k \geqslant \frac{\rho}{N_4 \| p_k \|} \geqslant \frac{\rho m_0}{N_4 N_2} \tag{9.22}$$

si $(N_3^2 + 1) \| z_{k+1} - z_k \|^2 = \alpha_k^2 N_4^2 \| p_k \|^2 \geqslant \rho^2$.

Il en résulte qu'on peut prendre pour constante une constante quelconque égale au plus à $\rho m_0/(N_4 N_2)$.

Développons $\varphi(z)$ en série de Taylor; nous établissons que dans une sphère S_{x_k} de rayon ρ

$$\varphi(z_{k+1})-\varphi(z_k)=\alpha_k(\varphi'(z_k),\ p_k)+\frac{\alpha_k^2}{2}(\varphi''(z_{kc})\,p_k,\ p_k)\leqslant$$

$$\leqslant\alpha_k(\varphi'(z_k),p_k)\left(1+\frac{\alpha_k}{2}\frac{N_2\|p_k\|^2}{(\varphi'(z_k),\ p_k)}\right).$$

Compte tenu de (9.18), la formule (9.19) entraîne

$$(\psi_k''p_k,\ p_k)=-(\varphi'(z_k),\ p_k)\geqslant m_0\|p_k\|^2.$$

Par suite,

$$\varphi(z_{k+1})-\varphi(z_k)\leqslant\alpha_k(\varphi'(z_k),\ p_k)\left(1-\frac{\alpha_k}{2}\frac{N_2}{m_0}\right).$$

On établit moyennant cette estimation et (9.22) que l'inégalité a manifestement lieu pour

$$\alpha_k=\min\left\{\delta,\ \frac{2m_0(1-\varepsilon)}{N_2}\right\},$$

ce qui veut dire, vu la borne inférieure de $f_0(x)$,

$$(\varphi'(z_k),\ p_k)\to 0. \tag{9.23}$$

Puisque $-(\varphi'(z_k),\ p_k)=((\psi_k'')^{-1}\varphi'(z_k),\ \varphi'(z_k))\geqslant m_0\|\varphi'(z_k)\|^2$ la convergence (9.23) implique $\|\varphi'(z_k)\|\to 0$, ce qui achève la démonstration du théorème.

Quiconque s'adresse à l'algorithme (9.20) doit se rappeler les remarques sur le choix des vecteurs z_k, y_k et du paramètre α_k que nous avons faites en étudiant la méthode (9.13).

Si la suite (9.20) converge vers la solution et $\varphi(z)$ à la minimisation de laquelle se réduit en fin de compte la résolution du problème de départ, vérifie la condition

$$\psi_k''\to\varphi''(z_k), \tag{9.24}$$

il y a la convergence superlinéaire. On l'établit en tenant compte de ce que sous cette condition (et vu que $(\psi_k''p_k,\ p_k)=-(\varphi'(z_k),\ p_k)$)

$$\varphi(z_{k+1})-\varphi(z_k)=\alpha_k(\varphi'(z_k),\ p_k)\left(1-\frac{\alpha_k}{2}\frac{(\varphi''(z_k)\,p_k,\ p_k)}{(\psi_k''p_k,\ p_k)}-\right.$$

$$\left.-\frac{\alpha_k}{2}\frac{((\varphi''(z_{kc})-\varphi''(z_k))\,p_k,\ p_k)}{(\psi_k''p_k,\ p_k)}\right)\to\alpha_k(\varphi'(z_k),\ p_k)\times$$

$$\times\left(1-\frac{\alpha_k}{2}-\frac{\alpha_k}{2}\frac{((\varphi''(z_{kc})-\varphi''(z_k))\,p_k,\ p_k)}{(\psi_k''p_k,\ p_k)}\right)$$

et la fonction $\varphi(z)$ est, par suite, de (9.18) et (9.24), fortement convexe dans un voisinage du minimum. En se rappelant ces circonstances, on démontre la convergence superlinéaire comme par exemple pour la méthode de Newton (§ 2, ch. II).

Dans certains problèmes, la méthode (9.20) a donc une convergence plus rapide que les méthodes du premier ordre. Mais elle peut s'avérer sensiblement plus coûteuse par itération vu la nécessité de calculer les dérivées secondes de $f_0(x)$.

4. Méthodes de minimisation de haute efficacité. Les méthodes de projection examinées aux numéros précédents sont dans un certain sens les analogues des méthodes du gradient et de Newton pour des problèmes d'extréma libres et elles connaissent les mêmes défauts que ces dernières : soit une convergence mauvaise (méthodes du premier ordre), soit le côut élevé de l'itération (celles du second ordre). Mais le fait que dans les algorithmes étudiés la résolution du problème initial se ramène à la recherche du minimum de fonctions sans contraintes (d'une ou de plusieurs selon qu'on est dans la condition (9.8) ou (9.12)) permet d'utiliser des algorithmes aussi efficaces que les méthodes des directions duales ou des directions conjuguées (§ 3-5, ch. II). Ainsi, pour des fonctions $f_i(x)$, $i = 1, \ldots, m$, telles qu'on ait la condition (9.8) et le système (9.1) admette pour tout z fixe une seule solution $y = y(z)$, si $\varphi(z) = f(z, y(z))$ est fortement convexe deux fois continûment dérivable, la convergence de toute méthode des directions duales ou conjuguées (destinée à la minimisation de $\varphi(z)$) vers la solution est superlinéaire. Si l'on utilise les versions avec reconstitution des matrices A_k^{-1} et H_k après un nombre fini de pas (voir n° 2, § 5, ch. II), la convergence est garantie sous les mêmes hypothèses sur $\varphi(z)$ que dans les méthodes du gradient.

Prenons à titre d'exemple le problème de minimisation d'une fonction quadratique $f_0(x)$ avec les contraintes linéaires $g(x) \equiv \equiv Ax + b = 0$, où $A = (a_{ij})$ est $m \times n$, $b = (b^1, \ldots, b^m)$.

Soit $|(a_{il})| \neq 0$, $i, l = 1, \ldots, m$. On pose alors $y = (x^1, \ldots \ldots, x^m)$, $z = (x^{m+1}, \ldots, x^n)$.

La fonction $y(z)$ étant linéaire, $\varphi(z)$ est une fonction quadratique de la variable z et elle est strictement convexe s'il en est de même de la fonction initiale $f_0(x)$. On trouve le minimum de $\varphi(z)$ au bout de $n - m$ pas par toute méthode utilisant des directions duales ou des directions conjuguées.

Si $f_0(x)$ et $f_i(x)$, $i = 1, \ldots, m$, vérifient les exigences du théorème 9.2 ((9.8) étant remplacé par une condition plus faible (9.12)), les directions duales et les directions conjuguées permettent de minimiser toute fonction $\varphi(z)$ intervenant dans la résolution du problème. En d'autres termes, le vecteur p_k et le paramètre α_k se définissent dans les algorithmes du type (9.9) comme dans les méthodes des directions duales ou conjuguées et les vecteurs z_k et y_k sont choisis comme dans la méthode (9.13). Les algorithmes ainsi construits (et qui s'effectuent avec reconstitution de la matrice A_k^{-1} ou H_k) convergent dans les mêmes conditions que les méthodes de

projection du premier et du second ordre, mais ils l'emportent sur ces dernières quant à leur efficacité; en particulier, une élévation relativement petite du coût de l'intération par rapport aux méthodes du premier ordre conduit à la convergence superlinéaire.

Dans la pratique, c'est justement par les algorithmes du type décrit dans le présent numéro qu'on doit résoudre le problème étudié.

On réduit le coût par itération (de la méthode (9.20)) en utilisant au lieu de $f_0''(x_k)$ la matrice D_k définie par le système

$$D_k(x_{k-i} - x_{k-i-1}) = f_0'(x_{k-1}) - f_0'(x_{k-i-1}), \quad i = 0, 1, \ldots, n-1$$

(l'analogue du système (II.3.6) permettant de construire les méthodes des directions duales), et en construisant le vecteur $p_k = -F_k^{-1}\varphi'(z_k)$, où $F_k = D_{kzz} + y'^*D_{kyy}y' + 2y'^*D_{kzy}$, les matrices D_{kzz}, D_{kyy}, D_{kzy} étant les composantes de D_k correspondant aux matrices f_{0zz}, f_{0yy}, f_{0zy}.

5. Sur la résolution du problème général de programmation mathématique. Soit à minimiser $f_0(x)$ dans les contraintes

$$f_i(x) \leqslant 0, \quad i = 1, \ldots, m, \tag{9.25}$$

qu'on peut ramener de diverses façons aux contraintes égalité. Introduisons par exemple les variables supplémentaires $x^{n+1}, \ldots, x^{n+m}$. Les contraintes (9.25) sont alors vérifiées pour les mêmes valeurs des variables $x^1, \ldots, x^n$ qui satisfont aux égalités

$$(x^{n+i})^2 + f_i(x) = 0, \quad i = 1, \ldots, m. \tag{9.26}$$

Le minimum de $f_0(x)$ dans les contraintes (9.25) coïncide donc avec le minimum de cette fonction dans les contraintes (9.26). On recherche le minimum de $f_0(x)$ sous les contraintes (9.26) à l'aide des méthodes du premier ordre examinées au n° 2.

Il est impossible de résoudre le dernier problème par la méthode (9.20) parce que $f_0(x)$ n'est pas strictement convexe dans l'espace E^{n+m}: on constate aisément la singularité dans E^{n+m} de la matrice $f_0''(x)$. C'est également la raison de l'impuissance des directions duales et conjuguées.

6. Remarques terminales. De toute la classe de méthodes de projection avec reconstitution des liaisons nous n'avons analysé que les algorithmes utilisant les formules (9.9). Pour certains problèmes cette mise en œuvre de la troisième (dernière) étape de l'itération peut s'avérer peu commode, ce qui appelle une autre tactique. On définit par exemple le point $x_{k+1} \in S_g$ de sorte que la quantité $\| x_{k+1}(\alpha) - \hat{x}_k(\alpha) \|$ minimise la distance entre $\hat{x}_k(\alpha)$ et l'ensemble S_g.

Le coût par itération des méthodes de projection diminue à mesure qu'on s'approche de la solution (ou d'un point stationnaire de

$f_0(x)$ sur S_g) puisque le problème de définition de x_{k+1} (moyennant $\hat{x}_k$) se simplifie progressivement. Ainsi, l'effort de calcul qu'exige la résolution du système (9.1) pour z_{k+1} fixe (i.e. le calcul de la fonction $y(z_{k+1})$) diminue d'une approximation à l'autre car le point $y_1(z_{k+1})$ approche toujours mieux la solution $y(z_{k+1})$ avec la croissance de k. En ce sens les méthodes de projection ont un grand avantage sur les techniques utilisant les fonctions de pénalisation dans lesquelles chaque nouvelle approximation exige qu'on résolve un problème plus ardu.

COMMENTAIRE

§ 1. La méthode de résolution des problèmes de programmation quadratique s'inspire de l'idée du gradient conjugué. Elle est d'un emploi particulièrement simple et commode dans le cas de contraintes simples sur les variables. Les problèmes de programmation quadratique se prêtent à de nombreuses autres méthodes dont certaines convergent en n pas (voir les monographies de *H. R. Künzi et W. Crelle* [83], *S. Zoukhovitski et L. Avdéeva* [14], *G. Zoutendijk* [104; 1], *V. Démianov et V. Malozémov* [9; 2]).

V. Ivanov [16; 2, 3], *V. Ivanov et V. Troutègne* [17] analysent l'efficacité et la précision de divers algorithmes.

§ 2. Notre exposé de la méthode des directions admissibles suit pour l'essentiel *S. Zoukhovitski, P. Poliak et M. Primak* [15; 1, 2]. La méthode du § 2 diffère de la méthode classique par la règle gouvernant le choix de la longueur du pas à chaque itération. Diverses versions des directions admissibles sont étudiées et justifiées en détail par *G. Zoutendijk* [104; 1, 2] et *D.M. Topkis et A. F. Veinott* [100].

§ 3. Pour le premier exposé de la méthode du gradient conditionnel voir *M. Frank et P. Wolfe* [63]. Elle a été ensuite étudiée par *V. Démianov et A. Roubinov* [10], *E. Lévitine et B. Poliak* [22], où on trouve les estimations de la vitesse de convergence. *M. D. Cannon et C. D. Cullum* [47] en ont prouvé l'exactitude.

E. Lévitine et B. Poliak [22] ont généralisé la méthode de Newton à des problèmes avec contraintes. La méthode de Newton à pas variable a été étudiée entre autres par *Y. Daniline* [7; 1, 2].

§ 4. En présentant la méthode du plan sécant les auteurs se sont inspirés de l'ouvrage de *J. E. Kelley* [81]. Ses diverses généralisations et l'estimation de la vitesse de convergence ($|f(x_k) - f(x_*)| \leqslant C/n$) pour une fonction $f(x)$ fortement convexe sont données dans l'article de *E. Lévitine et B. Poliak* [22].

§§ 5-6. L'exposé de la méthode de linéarisation suit le travail de *B. Pchénitchny* [29; 4] qui en démontre la convergence. Des résultats plus fins tels que la convergence en n pas pour la programmation linéaire, l'estimation locale de la vitesse de convergence et la convergence quadratique dans des cas spéciaux sont publiés pour la première fois. Il en est de même de l'application de cette méthode à un problème de minimax. Ce dernier problème a été étudié par *V. Démianov et V. Malozémov* [9; 2, 1] qui ont construit plusieurs algorithmes de descente correspondants.

Notons de plus que le problème de minimax peut être abordé par l'algorithme de gradient généralisé et ses variantes élaborés par N. Chor [36; 1-4].

§ 7. Bien que proposé pour la première fois, le procédé d'accélération de la convergence (n°s 1-3) est très proche quant à son idée de celui de *Y. Daniline et B. Pchénitchny* [8; 1, 2]. Les méthodes de résolution de systèmes d'équations sans calcul de dérivées des premiers membres ont intéressé de nombreux auteurs dont *F. J. Zeleznik* [103], *J. Barnes* [42], *C. G. Broyden* [46; 1], *V. Chamanski* [35].

Le n° 4 applique la méthode de Newton aux systèmes d'équations aux équations qui interviennent lorsqu'on formule les conditions nécessaires d'extrémum. Pour une étude plus poussée des questions correspondantes consulter *B. Poliak* [28; 3].

§ 8. Technique très employée, la méthode de pénalisation fait l'objet de nombreux ouvrages. Ses diverses propriétés sont traitées dans les monographies de *J. Céa* [50], *A. V. Fiacco et G. P. McCormick* [59], *E. Polak* [93; 2], dans les articles de *W. I. Zangwill* [102; 2], *E. Lévitine et B. Poliak* [22], *A. V. Fiacco* [58], et les estimations de sa vitesse de convergence dans *I. Erémine* [12; 1] et *D. Laenberger* [85]. La méthode des centres qui est étroitement liée avec la méthode de pénalisation a été proposée par *P. Huard* [77]. On trouve une variante nouvelle, peu étudiée, de la dernière technique dans *M. R. Hestenes* [72; 1].

§ 9. C'est *J. B. Rosen* [96] qui a été l'un des premiers à utiliser les méthodes de projection pour résoudre des problèmes avec contraintes. L'article de *A. Miele, H. Y. Huang et J. C. Heideman* [88] analyse des algorithmes de projection avec reconstitution des liaisons dans lesquelles la direction de déplacement est définie par une méthode du gradient ou par le gradient conjugué et on minimise à l'étape finale la distance du point aux liaisons.

Les algorithmes de projection exposés dans ce paragraphe ont été étudiés par *Y. Daniliné* [7; 3].

Parmi les procédés négligés par les auteurs citons la méthode des approximations de Fejer développée par *I. Erémine* [12; 2, 3] et la méthode du gradient généralisé de *N. Chor* [36; 1-4]. Pour des méthodes combinées destinées à la recherc e de l'extrémum voir *V. Ivanov* [16; 1]. On doit à *V. Karmanov* [29] de nombreux résultats fort intéressants sur la convergence des algorithmes de minimisation.

Quant aux ouvrages de synthèse consacrés aux méthodes numériques, citons *Y. Ermoliev* [13], *G. Zoutendijk* [104; 2], ainsi que *H. P. Künzi et W. Oettli* [84] et *F. Vassiliev* [4], où l'on trouve une vaste bibliographie sur la question considérée.

ANNEXE

SCHÉMAS DE CALCUL DE PRINCIPAUX ALGORITHMES

I. Méthode des directions duales (ch. II, § 3)

La méthode est destinée à minimiser une fonction convexe $f(x)$, $x \in E^n$.

Schéma de l'itération.

Soient x_0 un point quelconque et $s_{0,0}, s_{0,-1}, \ldots, s_{0,-n+1}$ un système arbitraire de vecteurs linéairement indépendants.

Pour $0 \leqslant k \leqslant n-1$ l'itération comprend les étapes suivantes:

1) on construit le point

$$x_{k+1} = x_k - \alpha_k f'(x_k), \tag{1}$$

avec α_k choisi par n'importe quel procédé du § 1, ch. II;

2) on pose

$$\begin{aligned} r_{k-1} &= x_{k+1} - x_k, \\ e_{k+1} &= f'(x_{k+1}) - f'(x_k); \end{aligned} \tag{2}$$

3) on calcule

$$(s_{k,\,k+n+1},\ e_{k+1}).$$

Si

$$|(s_{k,\,k-n+1},\ e_{k+1})| \geqslant \gamma \,\|s_{k,\,k-n+1}\|\,\|e_{k+1}\|, \tag{3}$$

où $\gamma > 0$ est une constante aussi petite qu'on le veut, alors on passe à 5).

Si

$$|(s_{k,\,k-n+1},\ e_{k+1})| < \gamma \,\|s_{k,\,k-n+1}\|\,\|e_{k+1}\|, \tag{4}$$

on passe à 4);

4) on pose

$$r_{k+1} = \beta_{k+1} \,\|s_{k,\,k-n+1}\|, \tag{5}$$

avec $\beta_{k+1} > 0$ choisi de façon qu'on ait la condition $\|r_{k+1}\| < \|r_k\|$.

On calcule le gradient $f'(x_k + r_{k+1})$ et on construit le vecteur $e_{k+1} = f'(x_k + r_{k+1}) - f'(x_k)$, puis on passe à 5);

5) on construit le système de vecteurs

$$\begin{aligned} s_{k+1,\,k+1} &= \frac{s_{k,\,k-n+1}}{(s_{k,\,k-n+1},\ e_{k+1})}, \\ s_{k+1,\,k-j} &= s_{k,\,k-j} - (s_{k,\,k-j},\ e_{k+1})\, s_{k+1,\,k+1}, \\ & j = 0, 1, \ldots, n-2. \end{aligned} \tag{6}$$

C'est l'étape finale de l'itération.

Pour $k \geqslant n$ on procède comme suit :

1) on construit le vecteur

$$p_k = -\sum_{i=0}^{n-1} (f'(x_k),\ s_{k,\,k-i})\, r_{k-i}\,;$$

2) on calcule la quantité $(f'(x_k),\ p_k)$.

Si $(f'(x_k),\ p_k) \lesseqgtr 0$, on construit le point x_{k+1} selon l'une des formules $x_{k+1} = x_k \pm \alpha_k p_k$, avec α_k choisi à partir de la condition (2.2), ch. II.

Si $(f'(x_k),\ p_k) = 0$, alors x_{k+1} est obtenu par la méthode du gradient (voir n° 1) pour $k \leqslant n-1$ (le processus itératif ultérieur s'effectue comme pour $k \leqslant n-1$ (étapes 2) — 5))).

REMARQUES.

1. Nous n'avons décrit qu'un seul schéma de calcul des méthodes des directions duales. Les premières itérations ($k \leqslant n-1$) obéissent ici à la méthode du gradient. Comme celle-ci garantit en général au début une décroissance suffisamment rapide de la fonction, un tel démarrage s'avère commode pour nombre de problèmes.

2. Les désignations des vecteurs de la base duale sont changées pour la raison de commodité (cf. (6) et (II.3.21)).

3. Si le vecteur r_{k+1} est choisi sous forme (5) et la fonction $f(x)$ est régulière fortement convexe (i.e. on est dans les conditions (II.2.4)), l'inégalité (3) a automatiquement lieu à condition de prendre la constante γ suffisamment petite. En effet, si $f(x)$ remplit les exigences formulées, alors $\| e_k \| \leqslant M \| r_k \|$ et on a l'estimation (II.5.18). Par suite

$$(s_{k,\,k-n+1},\ e_{k+1}) = \frac{1}{\beta_k}(r_{k+1},\ e_{k+1}) \geqslant$$

$$\geqslant \frac{m}{\beta_k} \| r_{k+1} \|^2 \geqslant \frac{m}{\beta_k M} \| r_{k+1} \| \, \| e_{k+1} \| = \frac{m}{M} \| s_{k,\,k-n+1} \| \, \| e_{k+1} \|.$$

Ainsi, l'inégalité (3) est valide quand $\gamma \leqslant \frac{m}{M}$.

4. L'expérience montre que γ peut être choisi très petit : $\gamma = 10^{-6} \div 10^{-15}$. Si la condition (3) n'est pas vérifiée même pour r_{k+1} sous forme (5), cela atteste que la matrice $f''(x)$ devient pour $x \to x_*$ de plus en plus mal conditionnée, i. e. la fonction à minimiser n'est pas fortement convexe. En particulier, les surfaces de niveau d'une telle fonction peuvent présenter des « vallées » bien prononcées. Il est alors impossible d'approcher très bien la solution par rapport à l'argument. Dans la pratique, on arrive cependant à obtenir des valeurs suffisamment voisines du minimum même dans le cas des fonctions non convexes possédant des surfaces de niveau décrites.

II. Méthode du gradient conjugué (ch. II, § 4)

La méthode est destinée à minimiser une fonction convexe $f(x)$, $x \in E^n$.

Schéma de l'itération.

Soient x_0 un point quelconque et $p_0 = -f'(x_0)$.

1) Si $\begin{cases} 0 \leqslant k \leqslant n-1, & \text{on passe à 2),} \\ k = n & \text{on passe à 5)}\,; \end{cases}$

2) on construit le point

$$x_{k+1} = x_k + \alpha_k p_k,$$

le facteur α_k étant défini par la condition

$$f(x_k + \alpha_k p_k) = \min_{\alpha \geqslant 0} f(x_k + \alpha p_k);$$

3) On calcule le vecteur

$$p_{k+1} = -f'(x_{k+1}) + \beta_{k+1} p_{k+1},$$

où

$$\beta_{k+1} = -\frac{(f'(x_{k+1}),\ f'(x_{k+1}) - f'(x_k))}{(f'(x_k),\ p_k)};$$

4) on passe à 1) ;

5) on pose $x_n = x_0$, $p_0 = -f'(x_n)$ et on recommence le processus (passage à 1)).

REMARQUE. Le coefficient β_{k+1} se définit par toute formule (II.4.73).

III. Méthode des directions admissibles (ch. III, § 2)

La méthode est destinée à résoudre le problème de programmation convexe : minimiser $f_0(x)$ avec les contraintes

$$f_i(x) \leqslant 0, \quad i = 1, \ldots, m,$$
$$Ax - b = 0,$$

où $x \in E^n$, $f_i(x)$, $i = 0, \ldots, m$, sont des fonctions convexes continûment dérivables, A est une matrice $l \times n$ et b un l-vecteur.

Notations :

$$I_{\bar{\delta}}(x) = \{i : f_i(x) \geqslant -\delta, \quad i = 1, \ldots, m\},$$
$$\|p\| = \max_{1 \leqslant j \leqslant n} |p^j|,$$

où $p \in E^n$, p^j étant les composantes du vecteur p.

Données de départ : x_0 est une approximation initiale vérifiant toutes les contraintes ; $\delta_0 > 0$, $\xi_i > 0$, $i = 0, \ldots, m$, sont des nombres positifs (en général quelconques).

Pas général de l'algorithme.

Le point x_k et le nombre $\delta_k > 0$ sont calculés.

1) Résoudre le problème de programmation linéaire

$$\min \eta$$
$$(f_i'(x_k),\ p) \leqslant \xi_i \eta, \quad i \in I_{\bar{\delta}_k}(x_k) \cup \{0\},$$
$$Ap = 0,$$
$$-1 \leqslant p^j \leqslant +1, \quad j = 1, \ldots, n.$$

Sa solution est η_k, p_k.

2) Si $\eta_k < -\delta_k$, alors

$$x_{k+1} = x_k + \alpha_k p_k, \quad \delta_{k+1} = \delta_k,$$

où $\alpha_k = \dfrac{1}{2^{q_0}}$ et q_0 est le premier entier $q = 0, 1, \ldots,$ qui vérifie les inégalités

$$f_0\left(x_k + \frac{1}{2^q} p_k\right) \leqslant f_0(x_k) + \frac{1}{2}\frac{1}{2^q}\eta_k,$$

$$f_i\left(x_k + \frac{1}{2^q} p_k\right) \leqslant 0, \quad i = 1, \ldots, m.$$

3) Si $\eta_k \geqslant -\delta_k$, alors $x_{k+1} = x_k$, $\delta_{k+1} = \dfrac{1}{2}\delta_k$.

4) Recommencer avec 1).

REMARQUE. Le choix des δ_0, ξ_i peut influer sur le processus, mais il doit être suggéré par l'analyse d'un problème concret. L'algorithme s'applique également à des problèmes non convexes.

V. Méthode de linéarisation (ch. III, § 5)

La méthode est utilisée pour résoudre le problème suivant : minimiser $f_0(x)$ sous les contraintes

$$f_i(x) \leqslant 0, \quad i = 1, \ldots, m,$$
$$f_i(x) = 0, \quad i = m+1, \ldots, m+l,$$

avec $f_i(x)$ des fonctions continûment dérivables.

Notations :

$$F(x) = \max\{0, f_1(x), \ldots, f_m(x), |f_{m+1}(x)|, \ldots, |f_{m+l}(x)|\},$$
$$I_\delta^-(x) = \{i : f_i(x) \geqslant F(x) - \delta, \ i = 1, \ldots, m\},$$
$$I_\delta^0(x) = \{i : |f_i(x)| \geqslant F(x) - \delta, \quad i = m+1, \ldots, m+l\},$$
$$\Phi_N(x) = f_0(x) + NF(x),$$
$$\|p\|^2 = \sum_{j=1}^{n} (p^j)^2.$$

Données de départ : x_0 est une approximation initiale arbitraire, N_0 un nombre suffisamment grand, $\delta_0 > 0$, $0 < \varepsilon < 1$.

Pas général de l'algorithme.

Le point x_k et les nombres N_k et δ_k sont construits.

1) On résout le problème

$$\min (f_0'(x_k), p) + \frac{1}{2}\|p\|^2,$$
$$(f_i'(x_k), p) + f_i(x_k) \leqslant 0, \quad i \in I_{\delta_k}^-(x_k),$$
$$(f_i'(x_k), p) + f_i(x_k) = 0, \quad i \in I_{\delta_k}^0(x_k).$$

Sa solution est p_k. Si le problème est incompatible, on pose $x_{k+1} = x_k$, $\delta_{k+1} = \frac{1}{2}\delta_k$, $N_{k+1} = N_k$ et on recommence avec 1).

2) Si le problème est compatible et p_k obtenu, on pose

$$x_{k+1} = x_k + \alpha_k p_k,$$
$$\delta_{k+1} = \delta_k,$$

où α_k est choisi égal à $\frac{1}{2^{q_0}}$ et q_0 est le premier des entiers $q = 0, 1, \ldots$, qui vérifie la relation

$$\Phi_{N_k}\left(x_k + \frac{1}{2^q} p_k\right) \leqslant \Phi_{N_k}(x_k) - \frac{1}{2^q}\varepsilon \|p_k\|^2.$$

3) Soient u_k^i, $i \in I_{\delta_k}^-(x_k) \cup I_{\delta_k}^0(x_k)$ les multiplicateurs de Lagrange du problème auxiliaire résolu à l'étape 1). Dans ce cas, on a $N_{k+1} = N_k$ si

$$N_k > \sum_{i \in I_{\delta_k}^-(x_k)} u_k^i + \sum_{i \in I_{\delta_k}^0(x_k)} |u_k^i|.$$

Dans le cas contraire,

$$N_{k+1} = 2\left[\sum_{i \in I^{-}_{\delta_k}(x_k)} u_k^i + \sum_{i \in I^{0}_{\delta_k}(x_k)} |u_k^i|\right].$$

4) On recommence avec 1).

REMARQUE. A partir d'un pas les nombres δ_k et N_k ne varient plus. L'algorithme exige un programme standard efficace de résolution du problème de programmation quadratique.

V. Algorithme de résolution d'un système d'équations sans calcul de dérivées (ch. III, § 6)

L'algorithme sert à résoudre le système d'équations

$$p(x) = 0,$$

où $x \in E^n$, $p(x)$ est une fonction vectorielle n-dimensionnelle de composantes dérivables $p^j(x)$, $j = 1, \ldots, n$.

Données de départ : les approximations initiales $x_1, \ldots, x_n$ sont choisies quelconques dans un voisinage suffisamment restreint de la solution. En particulier, tous les x_k, $k = 1, \ldots, n$, peuvent coïncider.

Notations : $\| p(x) \|^2 = \sum_{j=1}^{n} (p^j(x))^2 \cdot \varphi(k)$ vaut $1, 2, \ldots, n-1$ si la division de k par n donne respectivement $1, 2, \ldots, n-1$ pour reste ; $\varphi(k) = n$ si k est divisible par n.

Pas général de l'algorithme.

$x_1, \ldots, x_k$ sont déjà construits.

1) Résoudre en β_i, $i = 1, \ldots, n$, le système d'équations

$$\sum_{i=1}^{n} z_{k-n+i}\beta_i = -p(x_k),$$

où

$$z_j = \frac{1}{\| p(x_j) \|}\left[p(x_j + \| p(x_j) \| e_{\varphi(j)}) - p(x_j)\right],$$

e_i étant un vecteur de composantes nulles excepté la i-ième qui vaut 1.

2) Poser

$$x_{k+1} = x_k + \sum_{i=1}^{n} \beta_i e_{\varphi(k-n+i)}.$$

3) Recommencer avec 1).

BIBLIOGRAPHIE

[1] Берщанский Я. М.

Об одном методе решения задач линейного и выпуклого программирования, ЖВМ и МФ, 10, № 3 (1970), 621-629.

[2] Будак Б. М., Гольдман Н. Л.

О применении метода Ньютона к решению нелинейных краевых задач, Сб. «Вычислительные методы и программирование», вып. VI, изд-во МГУ (1967), 17-33.

[3] Вайнберг М. М.

1. Вариационный метод и метод монотонных операторов, «Наука», 1972.
2. Variational Methods for the Study of nonlinear Operators, Holden-day, San Francisco, 1964.

[4] Васильев Ф. П.

Лекции по методам решения экстремальных задач, изд-во МГУ (1974).

[5] Глазман И. М.

Релаксационные методы. В сб. «Труды 1-й зимней школы по математическому программированию» (Дрогобыч, 1968), т. I, М., 1969.

[6] Гольштейн Е. Г.

1. Двойственные задачи выпуклого и частично-выпуклого программирования в функциональных пространствах. В сб. «Исследования по математическому программированию», «Наука», 1968.
2. Выпуклое программирование. Элементы теории, «Наука», 1970.

[7] Данилин Ю. М.

1. Об одном подходе к задачам минимизации, ДАН СССР, 188, № 6 (1969) 1221-1222.
2. Методы минимизации, основанные на аппроксимации исходного функционала выпуклым, ЖВМ и МФ, 10, № 5 (1970), 1067-1080.
3. О минимизации функций в задачах с ограничениями типа равенств, Кибернетика, № 2 (1971), 88-95.
4. Методы сопряженных направлений для решения задач минимизации, Кибернетика, № 5 (1971), 122-136.

[8] Данилин Ю. М., Пшеничный Б. Н.

1. О методах минимизации с ускоренной сходимостью, ЖВМ и МФ, 10, № 6 (1970), 1341-1354.
2. Метод минимизации без вычисления производных, ЖВМ и МФ, 11, № 1 (1971), 12-21.

[9] Демьянов В. Ф., Малоземов В. Н.

1. К теории нелинейных минимаксных задач, Успехи мат. наук, XXVI, № 3 (1971), 53-104.
2. Введение в минимакс, «Наука», 1972.

[10] Демьянов В. Ф., Рубинов А. М.

Приближенные методы решения экстремальных задач, Изд. ЛГУ, 1968.

[11] Дубовицкий А. Я., Милютин А. А.

Задачи на экстремум при наличии ограничений, ЖВМ и МФ, 5, № 3 (1965), 395-453.

[12] Еремин И. И.

1. Метод штрафов в выпуклом программировании, Кибернетика, № 4 (1967), 63-67.

2. Метод Фейеровских приближений в выпуклом программировании, Мат. заметки, 3 (1968), 217-234.
3. Скорость сходимости метода Фейеровских приближений, Мат. заметки, 4 (1968), 53-61.

[13] Ермольев Ю. М.
Методы решения нелинейных экстремальных задач, Кибернетика, № 4 (1966), 1-17.

[14] Зуховицкий С. И., Авдеева Л. И.
Линейное и выпуклое программирование, «Наука», 1967.

[15] Зуховицкий С. И., Поляк Р. А., Примак М. Е.
1. Алгорифм для решения задачи выпуклого чебышевского приближения, ДАН СССР, 151, № 1 (1963), 27-30.
2. Алгорифм для решения задачи выпуклого программирования, ДАН СССР, 153, № 5 (1963), 991-994.

[16] Иванов В. В.
1. Теория приближенных методов и ее применение к численному решению сингулярных интегральных уравнений, «Наукова Думка», К., 1968.
2. Вопросы точности и эффективности вычислительных алгоритмов, Изд-во ИК АН УССР, Киев, 1969.
3. Об оптимальных алгоритмах минимизации функций некоторых классов, Кибернетика, № 4 (1972), 81-94.

[17] Иванов В. В., Трутень В. Е.
Анализ точности квадратичных программ, Кибернетика, № 4 (1969), 94-105.

[18] Исаев В. К., Сонин В. В.
Об одной модификации метода Ньютона численного решения краевых задач, ЖВМ и МФ, 3, № 6 (1963), 1114-1116.

[19] Канторович Л. В.
1. О методе наискорейшего спуска, ДАН СССР, 56 (1947), 233-236.
2. О методе Ньютона для функциональных уравнений, ДАН СССР, 60, 7 (1948).

[20] Карманов В. Г.
Оценки скорости сходимости итерационных методов минимизации, ЖВМ и МФ, 14, № 1 (1974).

[21] Лавров С. С.
Применение барицентрических координат для решения некоторых вычислительных задач, ЖВМ и МФ, 4, № 5 (1964), 905-911.

[22] Левитин Е. С., Поляк Б. Т.
Методы минимизации при наличии ограничений, ЖВМ и МФ, 6, № 5 (1966), 787-823.

[23] Любич Ю. И.
Наискорейший спуск. Труды второй зимней школы по матем. прогр. и смежн. вопр. Выпуск 1, Москва, 1969, 113-151.

[24] Любич Ю. И., Майстровский Г. Д.
Общая теория релаксационных процессов для выпуклых функционалов. Успехи мат. наук, 25, вып. 1, 1970.

[25] Майстровский Г. Д.
1. О сходимости метода сопряженных градиентов, ЖВМ и МФ, 11, № 5 (1971), 1291-1294.

2. Доказательство квадратичной сходимости метода сопряженных градиентов, Вычисл. матем. и выч. техника, Физ.-техн. инст. низких температур, Харьков, вып. 2 (1971), 3-5.

[26] Моисеев Н. Н.
Численные методы в теории оптимальных систем, «Наука», 1971.

[27] Обломская Л. Я.
Сравнение быстроты сходимости методов сопряженных градиентов и градиентного для квадратичных функционалов. В сборн. « Вопросы точности и эффективн. вычисл. алгор.», 4, Киев, 1968, 94-103.

[28] Поляк Б. Т.
1. Градиентные методы минимизации функционалов, ЖВМ и МФ, 3, № 4 (1963), 643-654.
2. Метод сопряженных градиентов. Труды второй зимней школы по матем. прогр. и смежн. вопр. Вып., 1, Москва, 1969, 152-201.
3. Итерационные методы, использующие множители Лагранжа для решения экстремальных задач с ограничениями типа равенств, ЖВМ и МФ, 10, № 5 (1970), 1098-1106.
4. О скорости сходимости метода штрафных функций, ЖВМ и МФ, 11, № 1 (1971), 3-11.

[29] Пшеничный Б. Н.
1. Необходимые условия экстремума, «Наука», 1969.
2. Принцип двойственности в задачах выпуклого программирования, ЖВМ и МФ, 5, № 1 (1965), 98-106.
3. Об одном алгоритме спуска, ЖВМ и МФ, 8, № 3 (1968), 649-652.
4. Алгоритмы для общей задачи математического программирования, Кибернетика, № 5 (1970), 120-125.
5. On the acceleration of convergence of aglorithms for solving optimal control. In « Computing methods in optimization problems 2 », Acad. Press., 1969.

[30] Пшеничный Б. Н., Ганжела И. Ф.
Алгоритм для решения задачи выпуклого программирования при линейных ограничениях, Кибернетика, № 3 (1970). 81-85.

[31] Смоляк С. А.
Квадратичная сходимость метода сопряженных градиентов. Труды III зимней школы по математическому программированию, МИСИ, 1970.

[32] Тихонов А. Н.
1. О регуляризации некорректно поставленных задач, ДАН СССР, 153, 1 (1963), 49-52.
2. Об устойчивости алгоритмов для решения вырожденных систем линейных алгебраических уравнений, ЖВМ и МФ, 5, № 4 (1965).

[33] Фихтенгольц Г. М.
Курс дифференциального и интегрального исчисления. 2, Физматгиз, 1959.

[34] Фридман В. М.
О сходимости методов наискорейшего спуска, Успехи мат. наук, 17, 3, 1962.

[35] Шаманский В. Е.
1. Методы численного решения краевых задач на ЭЦВМ. Часть II. «Наукова Думка», Киев, 1966.
2. О некоторых вычислительных схемах итерационных процессов, Успехи мат. наук. 14, № 1 (1962), 100-109.

[36] Ш о р Н. З.
1. Обобщенный градиентный спуск. Труды первой зимней школы по математическому программированию. г. Дрогобыч, М., 1969, 578-585.
2. О скорости сходимости обобщенного градиентного спуска, Кибернетика, № 3, (1968), 98-99.
3. Использование операции растяжения пространства в задачах минимизации выпуклых функций, Кибернетика, № 1 (1970), 6-12.
4. О скорости сходимости обобщенного градиентного метода с растяжением пространства, Кибернетика, № 2 (1970), 80-85.

[37] Я к о в л е в М. Н.
О некоторых методах решения нелинейных уравнений. Труды матем. ин-та АН СССР, 84 (1965), 8-40.

[38] A l t m a n M.
Generalized gradient methods of minimizing a functional, Bull. Acad. Polon. Sci., 14, N° 6 (1966), 313-318.

[39] A r r o w K. J., H u r w i c z L., U z a w a H.
Studies in linear and non-linear programming, Stanford Univ. Press, Stanford California, 1958.

[40] A u s l e n d e r A.
Méthodes et théorèmes de dualité, C. R. Acad. Sci. Paris, Ser. A 267 (1968), 114-117.

[41] B a l a k r i s h n a n A. V., N e u s t a d t L. W.
Computing Method in Optimization Problems, Academic Press, New York-London, 1964.

[42] B a r n e s J.
An algorithm for solving nonlinear equations based on the secant method, Computer J., 8 (1965), 66-72.

[43] B o u r b a k i N.
Eléments de mathématique, f. XV, Livre 5. Espaces vectoriels topologiques, Paris, 1966.

[44] B r e n t R. P.
Algorithms for finding zeros and extrema of fuctions without calculating derivatives, STAN-CS, 71-198, February, 1971.

[45] B r o w n K. M., D e n n i s J. E.
On Newton-like iteration functions : general convergence theorems and a specific algorithm, Numer. Math., 12 (1968).

[46] B r o y d e n C. G.
1. A class of methods for solving nonlinear simultaneaus equations, Math. Comput., v. 19 (1965), 577-593.
2. Quasi-Newton Methods and Their Application to Function Minimization, Math. Comput., 21, N° 99 (1967), 368-381.
3. The convergence of single-rank quasi-Newton methods, Math. Comput., 24 (1970).

[47] C a n n o n M. D., C u l l u m C. D.
A tight upper bound on the rate of convergence of the Frank-Wolfe algorithm, SIAM J. on Control, 6 (1968), 509-516.

[48] C a n n o n M. D., E a t o n J. H.
A new algorithm for a class of quadratic programming problems with application to control, SIAM J. on Control, 4, N° 1 (1966), 34-45.

[49] Cauchy A. L.
Méthode générale pour la résolution des systèmes d'équations simultanées, C. R. Acad. Sci. Paris, 25, 1847.

[50] Céa J.
Optimisation. Théorie et algorithmes, Dunod, Paris, 1971.

[51] Collatz L.
Funktionalanalysis und numerische Mathematik, Berlin [u.a], Springer, 1964.

[52] Daniel J. W.
1. The approximate minimization of functionals, PRENTICE-HALL, INC., 1971.
2. Convergence of the conjugate gradient method with computationally convinient modifications, Numer. Math., 10, N° 2 (1967), 125-131.
3. The conjugate gradient method for linear and nonlinear operator equations, SIAM J. Numer. Anal., 4 (1967), 10-26.

[53] Dantzig G. B.
1. Linear programming and extensions, Princeton University Press, Princeton, 1963.
2. Linear control processes and mathematical programming, SIAM J. on Control, 4, N° 1 (1966), 56-60.

[54] Davidon W. C.
1. Variable metric method for minimization, Argonne Nation. Labor., Rep. NANL-5990, 1959.
2. Variance algorithm for minimization, Computer J., 10, N° 4 (1968), 406-410.

[55] Dennis I. E.
On Newton-like methods, Numerische Mathematik, Bd. 11, Heft 4, 1968.

[56] Dunford N., Schwartz J.
Linear operators, New York-London, Interscience publ., 1958-1963.

[57] Faddeev D. K., Faddeeva V. N.
Computational methods of linear algebra, Freeman, San Francisco, 1963.

[58] Fiacco A. V.
Penalty method for mathematical programming in E^n with general constraint sets, J. Optim. Theory Applic., 6, N° 3 (1970), 252-268.

[59] Fiacco A. V., McCormick G. D.
Non Linear Programming : Sequential Unconstrained Minimization Technique, Wiley, 1968.

[60] Fletcher R.
1. A new approach to variable metric algorithms, Computer J.,13, N° 3 (1970), 317-322.
2. Function minimization without evaluating derivatives, A review, Comput. J., 8, N° 1 (1965), 33-41.

[61] Fletcher R., Powell M. J. D.
A rapidly convergent descent method for minimization, Computer J., 6, N° 2 (1963), 163-168.

[62] Fletcher R., Reeves C. M.
Function minimization by conjugate gradients, Computer J., 7, N° 2 (1964), 149-154.

[63] Frank M., Wolfe P.
An algorithm for quadratic programming, Naval Res. Log. Quart., 3 (1956), 95-110.

[64] Gale D.
The theory of linear economic models, New York [a, o.], McGraw-Hill, 1960.

[65] Gilbert E. G.
An iterative procedure for computing the minimum of a quadratic form on a convex set, SIAM J. on Control, 4, N° 1 (1966), 61-80.

[66] Goldfarb D.
A family of variable-metric methods derived by variational means, Math. Comput., 24, N° 109 (1970), 23-26.

[67] Goldstein A. A.
1. On steepest descent, J. SIAM Control, 3 (1965), 147-151.
2. Minimizing Functionals on Normed Linear Spaces, SIAM J. on Control, 4, N° 1 (1966), 81-89.

[68] Goldstein A. A., Price J. F.
An effective algorithm for minimization, Numer Math., 10, 3 (1967), 184-189.

[69] Greenstadt J.
Variations on variable-metric methods, Math. Comput., 24, N° 109 (1970), 1-22.

[70] Halkin E., Neustadt L. W.
General necessary conditions for optimization problems, Proc. Nat. Acad. Sciences, 56 (1956), 1066-1071.

[71] Hayes R. M.
Iterative methods of solving linear problems in Hilbert space. In « Contributions to the solution of systems of linear equations and determination of eigenvalues ». Ed. O. Taussky, Nat. Bur. Stand. Appl. Math., 39 (1954), 71-104.

[72] Hestenes M. R.
1. Multiplier and gradient methods, J. Optim. Theory Applic., 4, N° 5 (1969).
2. Calculus of variations and optimal control theory, Wiley and Sons, Inc., 1966.
3. The conjugate gradient method for solving linear systems, Proc. Symp. Appl. Math., 6 (1956) 86-102.

[73] Hestenes M. R., Stiefel E.
Methods of conjugate gradients for solving linear systems, J. Res. Nation. Bureau Stand., 49, N° 6 (1952), 409-436.

[74] Horwitz L. B., Sarachik P. E.
Davidon's method in Hilbert space, SIAM J. on Appl. Math., 16, 4 (1968).

[75] Huang H. Y.
Unifed approach to quadratically convergent algorithms for function minimization, J. Optim. Theory Applic., 5, N° 6 (1970).

[76] Huang H. Y., Levy A. V.
Numerical experiments on quadratically convergent algorithms for function minimization, J. Optim. Theory Applic., 6, N° 3 (1970), 269-282.

[77] Huard P.
Rezolution of mathematical programming with nonlinear constraints by the method of centers. In « Nonlinear Programming », Abadie ed. North Holland Publ., 1967, 208-219.

[78] Kantorovich L. V., Akilov G. P.
Functional Analysis in Normed Spaces, Mac Millan, 1964.

[79] Karlin S.
Mathematical Methods and Theory in Games, Programming and Economics, Addison-Wesley Reading Mass, 1959.

[80] Kelley H. J.
Method of Gradients. In « Optimization techniques. With applications to Aerospace Systems ». New-York, Acad. Press. London, 1962.

[81] Kelley J. E.
The cutting plane method for solving convex programs, J. Soc. Ind. Appl. Math., 8, N° 4 (1960), 703-712.

[82] Kolmogorov A., Fomine S.
Eléments de la théorie des fonctions et de l'analyse fonctionnelle, Ed. de Moscou, 1974.

[83] Künzi H. P., Crelle W.
Nichtlineare Programmierung, Berlin [u.a.], Springer, 1962.

[84] Künzi H. P., Oettli W.
Nichtlineare optimierung: neuere verfahren bibliographie, Springer-Verlag, Berlin-Heidelberg-New-York, 1969.

[85] Laenberger D.
Convergence rate of penalty-function scheme, J. Optim. Theory Applic., 7, N° 1 (1971), 39-51.

[86] Ljusternik L. A., Sobolew W. I.
Elemente der Functionalanalysis, Berlin Akad-Verlag, 1968.

[87] McCormick G. P., Pearson J. D.
Variable metric methods and unconstrained optimization, Confer. on Optimal., Keele Hall, England, March 1968.

[88] Miele A., Huang H. Y., Heideman J. C.
Sequential gradient-restoration algorithm for the minimization of constrained function — ordinary and conjugate gradient versions, J. Optim. Theory Applic., 4, N° 4 (1969).

[89] Murtagh B. A., Sargent R. W. H.
Computational experience with quadratically convergent minimization methods, Computer J., 13, N° 2 (1970), 185-194.

[90] Neustadt L. W.
An abstract variational theory with applications to broad class of optimization problems. I. General theory, SIAM J. on Control, 4, N° 3 (1966), 505-527.

[91] Ostrowski A.
Solution of equations and systems of equations, New-York-London, Acad. Press, 1966.

[92] Pearson J. D.
Variable metric methods of minimization, Computer J., 12, N° 2 (1969), 171-178.

[93] Polak E.
1. On primal and dual methods of solving discrete optimal control problems. In « Computing methods in optimization problems. 2 », Acad. Press, New-York-London, 1969.

BeatriceGloria_personal library

2. Computational methods in optimization : a unifed approach, Acad. Press, New-York, 1971.

[94] Powell M. J. D.
1. A survey of numerical methods for unconstrained optimization. SIAM Rev., 2, N° 1 (1970) 79-97.
2. An efficient method for finding the minimum of a function of several variables without calculating derivatives, Computer J., 7, N° 2 (1964), 155-162.
3. On the convergence of the variable metric algorithm, Report T. P. 382, AERE, Harwell, England, 1970.

[95] Rockafellar R. T.
Convex analysis, Princeton (N. Y.), Princeton Univ. Press, 1970.

[96] Rosen J. B.
The gradient projection method for nonlinear programming. Part I : Linear constraints ; Part II : Nonlinear constraints. SIAM J. on Appl. Math., 8 (1960), 181-217 ; 9 (1961), 514-532.

[97] Smith C. S.
The automatic computation of maximum likelihood estimates, N.C.B. Sci. Dept. Report SC 846 (MR) 40.

[98] Sorensen H. W.
Comparison of some conjugate direction procedures for function minimization, J. Franklin Institute, 288, 421 (1969).

[99] Tokumaru H., Adachi N., Goto K.
Davidon's method for minimization problems in Hilbert space with an application to control problems, SIAM J. on Control, 8, N° 2 (1970).

[100] Topkis D. M., Veinott A. F.
On the convergence of some feasible direction algorithms for nonlinear programming, SIAM J. on Control, 5, N° 2 (1967), 268-279.

[101] Warga J.
A convergent procedure for convex programming, J. Soc. Ind. Appl. Math., 11, N° 3 (1963), 579-587.

[102] Zangwill W. I.
1. Minimizing a function without calculating derivatives, Computer J., 10, N° 3 (1967), 293-296.
2. Nonlinear programming via penalty functions, Management Sci., N° 13 (1967), 344-368.

[103] Zeleznik F. J.
Quasi-Newton methods for nonlinear equations, J. Assoc. Comput., Mach., 15, N° 2 (1968), 265-271.

[104] Zoutendijk G.
1. Methods of feasible directions, Amsterdam [a.o.], Elsevier, 1960.
2. Nonlinear programming : a numerical survey, SIAM J. on Control, 4, N° 1 (1966), 194-210.

BIBLIOTHÈQUE DE LA

TABLE DES MATIÈRES

Préface 5

Chapitre premier. INTRODUCTION À LA THÉORIE DE LA PROGRAMMATION MATHÉMATIQUE 7

§ 1. Ensembles convexes 7
1. Définition. Théorème de séparation (7). 2. Cônes convexes (10). 3. Ensembles strictement et fortement convexes (13).

§ 2. Fonctions convexes 14
1. Définition. Propriétés fondamentales (14). 2. Propriétés de dérivabilité (15). 3. Fonctions strictement et fortement convexes (17). 4. Fonctions concaves (20).

§ 3. Programmation convexe 20
1. Position du problème. Propriétés fondamentales (20). 2. Conditions nécessaires de minimum (21). 3. Théorème de Kuhn et Tucker (24). 4. Problème dual (24). 5. Problème de programmation linéaire (25). 6. Problème de programmation quadratique (27).

§ 4. Conditions nécessaires de minimum 28
1. Définitions fondamentales (29). 2. Conditions nécessaires de minimum (29). 3. Problème de minimax (35). 4. Conditions nécessaires du deuxième ordre (36).

§ 5. Certains renseignements complémentaires 38

Commentaire 38

Chapitre II. MÉTHODES DE MINIMISATION DE FONCTIONS SANS CONTRAINTES 40

§ 1. Méthodes du gradient 41
1. Méthode de la plus grande pente (41). 2. Certaines variantes de la méthode de la plus grande pente (46). 3. D'autres méthodes du gradient (48). 4. Analyse qualitative des méthodes (51).

§ 2. Méthode de Newton à pas variable 53
1. Construction de la méthode (53). 2. Théorèmes sur les propriétés de la méthode (54). 3. Versions de la méthode de Newton généralisée (57). 4. Discussion des propriétés de la méthode de Newton (60).

§ 3. Méthodes des directions duales 61
1. Sur le choix du schéma des méthodes (61). 2. Justification des méthodes (63). 3. Construction de divers algorithmes (67). 4. Définition du vecteur p_k (69). 5. Démarrage du processus (72). 6. Minimisation d'une forme quadratique (73). 7. Discussion des propriétés des méthodes (73).

§ 4. Méthodes des directions conjuguées. Minimisation de fonctions quadratiques 75
1. Directions conjuguées et leurs propriétés (75). 2. Construction des méthodes (78). 3. Propriétés générales des méthodes (82). 4. Algorithmes concrets (85). 5. Minimisation d'une fonction quadratique convexe (90). 6. Discussion des résultats (93).

§ 5. Méthodes des directions conjuguées. Minimisation de fonctions quelconques 94
1. Applicabilité des méthodes (94). 2. Théorème sur la convergence des méthodes (95). 3. Etude des propriétés de divers algorithmes (101). 4. Etude de la vitesse de convergence (110). 5. Discussion des résultats (115).

§ 6. Méthodes directes 118
1. Préliminaires (118). 2. Construction des méthodes des directions

duales (119). 3. Quelques remarques sur la mise en œuvre des méthodes des directions duales (125). 4. Méthodes des directions conjuguées (127). 5. Discussion des résultats (132).

Commentaire 133

Chapitre III. MÉTHODES DE RÉSOLUTION DE PROBLÈMES AVEC CONTRAINTES 135

§ 1. Problème de programmation quadratique 135
1. Opérateurs de projection (136). 2. Minimisation d'une fonction quadratique sur un sous-espace (137). 3. Algorithme pour le problème général de la programmation quadratique (139). 4. Aspects numériques (146). 5. Problème de programmation quadratique avec des contraintes simples (148).

§ 2. Méthode des directions admissibles 149
1. Méthode du choix d'une direction admissible (150). 2. Algorithme de la méthode des directions admissibles (153). 3. Justification de la convergence de l'algorithme (154). 4. Construction de l'approximation initiale (157).

§ 3. Méthode du gradient conditionnel et méthode de Newton 158
1. Règle régissant le choix de la longueur du pas (158). 2. Description de l'algorithme (159). 3. Justification et évaluation de la convergence de l'algorithme (159). 4. Evaluation de la convergence dans le cas d'un domaine fortement convexe (163). 5. Méthode de Newton à pas variable (164). 6. Propriétés de la méthode de Newton (165).

§ 4. Méthode du plan sécant 171
1. Algorithme (172). 2. Aspects numériques (173). 3. Remarques terminales (174).

§ 5. Méthode de linéarisation 174
1. Hypothèses fondamentales (175). 2. Formulation de l'algorithme (176). 3. Convergence de l'algorithme (176). 4. Aspects numériques (180). 5. Certaines généralisations (183). 6. Problème de programmation linéaire (186). 7. Evaluation locale de la convergence (190).

§ 6. Méthode de linéarisation (résolution de systèmes d'égalités et d'inégalités et recherche du minimax) 196
1. Systèmes d'égalités et d'inégalités (196). 2. Convergence de l'algorithme (198). 3. Remarques (200). 4. Conditions suffisantes de convergence (201). 5. Résolution d'un problème de minimax (204).

§ 7. Accélération locale de la convergence 209
1. Position du problème. Formules fondamentales (210). 2. Algorithme (214). 3. Aspects numériques. Application à un problème de programmation mathématique (217). 4. Un problème de minimisation avec contraintes égalité (218).

§ 8. Méthode de pénalisation 219
1. Justification' de la méthode de pénalisation (220). 2. Programmation convexe (224). 3. Aspects numériques (226). 4. Méthode de Fiacco et McCormick (227).

§ 9. Méthodes de projection avec reconstitution des liaisons 228
1. Schéma de construction des méthodes (228). 2. Méthodes du premier ordre (232). 3. Une méthode du second ordre (235). 4. Méthodes de minimisation de haute efficacité (238). 5. Sur la résolution du problème général de programmation mathématique (239). 6. Remarques terminales (239)

Commentaire 240

Annexe. SCHÉMAS DE CALCUL DE PRINCIPAUX ALGORITHMES 242

Bibliographie 247

Imprimé en Union Soviétique